ITE102 – Artificial Intelligence
Final Project

Having gained a foundational understanding of key concepts in Artificial Intelligence, it is now time to explore various AI models designed for specialized tasks.

For your Final Examination, you will work with a selection of models available on Hugging Face, a platform where users can share and access open-source Machine Learning models and datasets.

Hugging Face hosts over 100,000 models and 200,000 datasets, with contributions from major tech companies such as Microsoft, Nvidia, Meta, Google, and OpenAI.

One significant area in machine learning is Computer Vision—the process of extracting meaningful information from images.

This project will focus on important Computer Vision tasks, including:

1. Depth Estimation
2. Image Classification
3. Object Detection
4. Text-to-Image
5. Text-to-Video

Follow the project guide, completing each step and addressing the questions as directed.

You will be using this Google Colab Notebook. Create a copy of the notebook and make any necessary modifications.

1. [25 pts] Depth Estimation

1. [1 pt] What is Depth Estimation?

2. [2 pts] What are the use cases of Depth Estimation?

Use cases	Description
1.	
2.	

3. [4 pts] Depth Anything V2 and MiDaS 3.0 are two models used for Depth Estimation. What are the differences between them?

	Depth Anything V2	MiDaS 3.0
Architecture	Depth Prediction Transformer (DPT)	
Backbone		

Content Curated by Pollux M. Rey

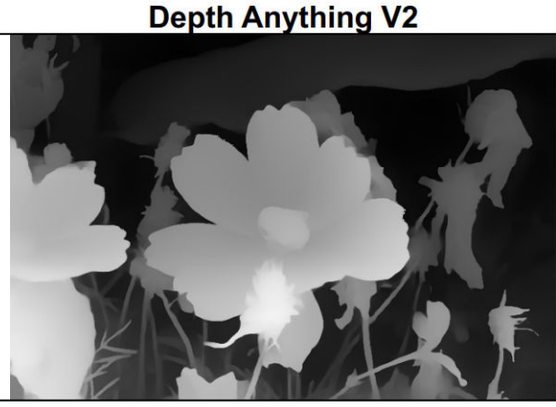
Amount of Training Data		

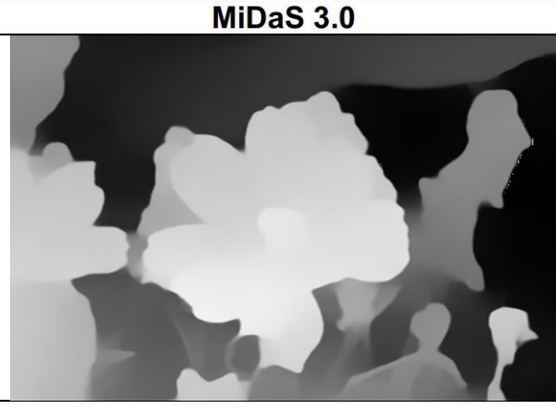
4. [3 pts] Find a **close-up photograph** online. Insert the image into the box below.

   Example image:

5. [10 pts] Use the photograph as input for both models to obtain the depth estimations. Place the resulting images into the corresponding boxes.

   Example images:

Depth Anything V2


MiDaS 3.0


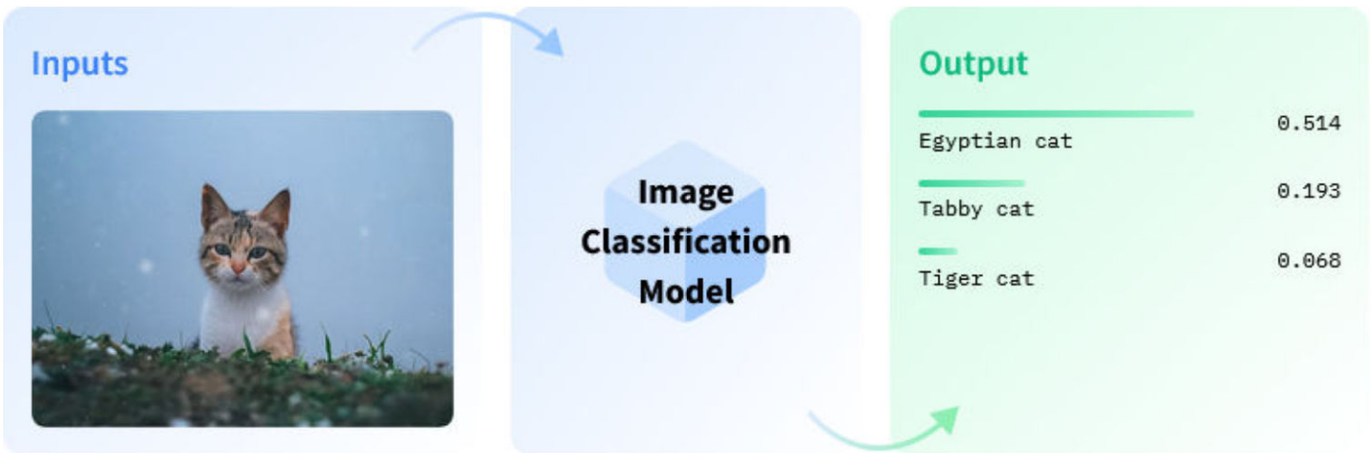
Depth Anything V2

Content Curated by Pollux M. Rey

MiDaS 3.0

6. [5 pts] What are your observations about the outputs? Which model do you think provided a better result, and why? Include a brief comparison.

II. [28 pts] Image Classification



1. [1 pt] What is Image Classification?

2. [2 pts] What are the use cases of Image Classification?

Use cases	Description
1.	
2.	

ResNet-50 is a model used for Image Classification. It has been pre-trained on the ImageNet-1k dataset.

3. [1 pt] What does "1k" in ImageNet-1k mean?

4. [1 pt] What image resolution was used to pre-train ResNet-50?

_____ x _____ pixels

Content Curated by Pollux M. Rey

5. [3 pts] Find **any photograph** online. Insert the image into the box below. Resize it to a height of 2 inches to keep it manageable, and center it within the box.

   Example image:

6. [5 pts] Use the photograph as input for ResNet-50 to classify the image. Place the output in the box below and highlight the label with the highest score.
   If the label with the highest score does not match the image, replace it.

   Example output:

```
[{'label': "jack-o'-lantern", 'score': 0.9999809265136719},
 {'label': 'four-poster', 'score': 6.093291062825301e-07},
 {'label': 'table lamp', 'score': 5.309839821165951e-07},
 {'label': 'stove', 'score': 3.11864340574175e-07},
 {'label': 'Yorkshire terrier', 'score': 2.2864626965201e-07}]
```

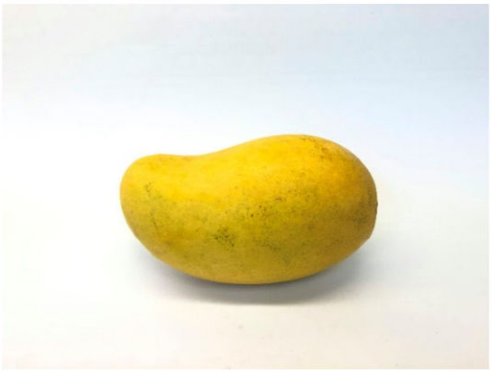
7. [2 pts] What do “label” and “score” in the output mean?

	Description
label	
score	

Content Curated by Pollux M. Rey

8. [3 pts] Find **any photograph** online in which the **label with the highest score does not match the image.**

   Example image:



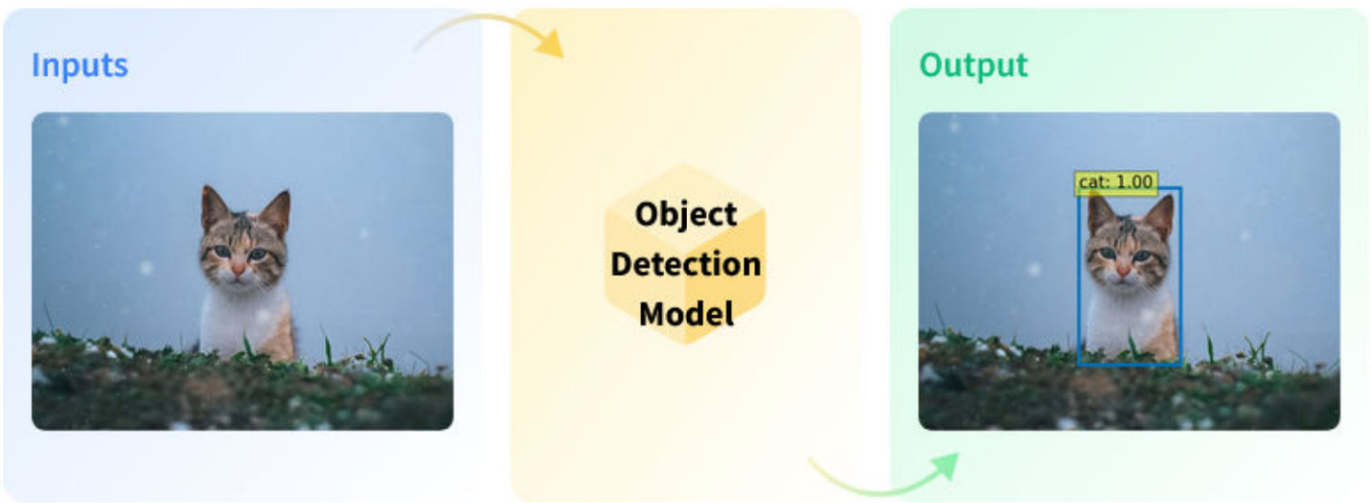
9. [5 pts] Place the output in the box below and highlight the label with the highest score.

   Example output:

```
[{'label': 'spaghetti squash', 'score': 0.5103492736816406},
 {'label': 'butternut squash', 'score': 0.4446260333061218},
 {'label': 'banana', 'score': 0.00175315304659307},
 {'label': 'orange', 'score': 0.0013020550832152367},
 {'label': 'acorn squash', 'score': 0.0010306020267307758}]
```

10. [5 pts] Why was it not labeled correctly?

III. [55 pts] Object Detection



Content Curated by Pollux M. Rey

1. [1 pt] What is Object Detection?

2. [4 pts] What are the use cases of Object Detection?

Use cases	Description
1.	
2.	
3.	
4.	

3. [3 pts] What is the difference between Image Classification and Object Detection?

4. [10 pts] DETR (End-to-End Object Detection) with ResNet-50 backbone and YOLOS are two models used for Object Detection. What are the differences between them?

	DETR	YOLOS
Type of Transformer		
Loss Calculation		
Training Data		
Training		
Average Precision on COCO 2017 Validation		

5. [3 pts] Find **any photograph with three or more objects** online. Insert the image into the box below. Resize it to a height of 2 inches to keep it manageable, and center it within the box.

Example image:

Content Curated by Pollux M. Rey

6. [10 pts] Use the photograph as input for both models to detect objects. Place the outputs in the boxes below.

Example outputs:

DETR

```
[{'score': 0.9976153373718262,
  'label': 'orange',
  'box': {'xmin': 350, 'ymin': 486, 'xmax': 488, 'ymax': 601}},
 {'score': 0.9169067144393921,
  'label': 'dining table',
  'box': {'xmin': 0, 'ymin': 533, 'xmax': 1198, 'ymax': 622}},
 {'score': 0.9954829216003418,
  'label': 'orange',
  'box': {'xmin': 628, 'ymin': 474, 'xmax': 770, 'ymax': 600}},
 {'score': 0.9949440360069275,
  'label': 'banana',
  'box': {'xmin': 665, 'ymin': 233, 'xmax': 1144, 'ymax': 578}}]
```

YOLOS

```
[{'score': 0.9911884069442749,
  'label': 'orange',
  'box': {'xmin': 341, 'ymin': 492, 'xmax': 494, 'ymax': 602}},
 {'score': 0.85792475938797,
  'label': 'apple',
  'box': {'xmin': 616, 'ymin': 479, 'xmax': 784, 'ymax': 607}},
 {'score': 0.9961064457893372,
  'label': 'banana',
  'box': {'xmin': 649, 'ymin': 257, 'xmax': 1155, 'ymax': 583}}]
```

DETR

YOLOS

7. [6 pts] What do the following items in the output mean?

	Description
label	
score	
xmin	
ymin	
xmax	
ymax	

Content Curated by Pollux M. Rey

8. To combine the output with the image, use the following code and run it.

```
from PIL import Image, ImageDraw, ImageFont
import matplotlib.pyplot as plt

image_copy = image.copy()
draw = ImageDraw.Draw(image_copy)

for item in output:
     box = item['box']
     label = item['label']
     score = item['score']

xmin, ymin, xmax, ymax = box['xmin'], box['ymin'], box['xmax'],
box['ymax']

draw.rectangle([(xmin, ymin), (xmax, ymax)], outline="red", width=2)

text = f"{label}: {score:.2f}"
draw.text((xmin, ymin - 40), text, fill="red", font_size=40)

plt.imshow(image_copy)
plt.axis('off')
plt.show()
```

[3 pts] What does the code do?

9. [10 pts] Place the resulting images produced by the code in the boxes below.

Example images:

DETR


YOLOS


Content Curated by Pollux M. Rey

DETR

YOLOS

10. [5 pts] What are your observations about the outputs? Which model provided a better result, and why?
Refer to the model comparison if one produced a better result than the other.

IV. [33 pts] Text-to-Image

1. [1 pt] What is Text-to-Image?

2. [4 pts] What are the use cases of Text-to-Image?

Use cases	Description
1.	
2.	
3.	

Content Curated by Pollux M. Rey

4.	

3. [10 pts] Stable Diffusion is a popular text-to-image model. Currently, the latest version is 3.5; however, for this project, we will use the earlier versions **1.5 and XL 1.0**. What are the differences among these versions?

	1.5	**XL 1.0**
Release Date		
Output Image Resolution	_____ x _____ pixels	_____ x _____ pixels
Text Conditioning Model		
Good Use Cases		
Poor Use Cases		

4. [3 pts] Provide a school-appropriate prompt for generating images with both versions.
To generate an image using Stable Diffusion on Google Colab, make sure to change runtime type to T4 GPU.

   Example: An astronaut riding a green horse

5. [10 pts] Place the resulting images into the corresponding boxes.

   Example images:

Stable Diffusion 1.5


Stable Diffusion XL 1.0

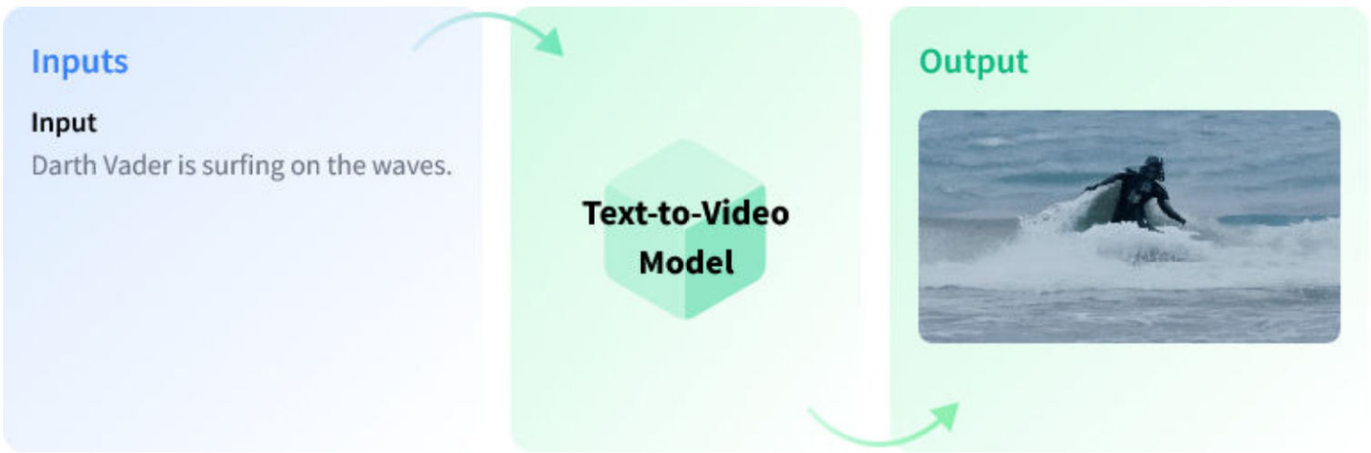

Content Curated by Pollux M. Rey

Stable Diffusion 1.5

Stable Diffusion XL 1.0

6. [5 pts] What are your observations about the outputs? Which model do you think provided a better result, and why? Include a brief comparison.

5. [36 pts] Text-to-Video



1. [1 pt] What is Text-to-Video?

2. [3 pts] What are the use cases of Text-to-Video?

Use cases	Description
1.	
2.	
3.	

The Text-to-Video Synthesis Model in Open Domain by ModelScore is a model used for video generation.

Content Curated by Pollux M. Rey

3. [3 pts] The Text-to-Video Synthesis Model utilizes three different models during the video generation process. What are these models?

1. 2. 3.

4. [1 pt] How many parameters are included in the model?

5. [3 pts] Provide a school-appropriate prompt for generating videos.
   To generate a video using Text-to-Video Synthesis Model on Google Colab, make sure to change runtime type to T4 GPU.

   Example: An astronaut riding a green horse

6. Create three videos using the same prompt, adjusting the settings as follows:
   1. num_inference_steps = 25
   2. num_inference_steps = 50
   3. num_inference_steps = 100

7. [5 pts] What are your observations about the outputs? Which settings do you think provided a better result, and why?

8. [15 pts] Upload the videos to Google Drive and place the links to these videos into the corresponding boxes.

num_inference_steps	Link
25	
50	
100	

9. [5 pts] Why do the generated videos have a Shutterstock watermark?

VI. [20 pts] Place your Google Colab notebook link here.

The outputs you created should be displayed in your Google Colab notebook.

Link:

Content Curated by Pollux M. Rey