Image and Video Analysis

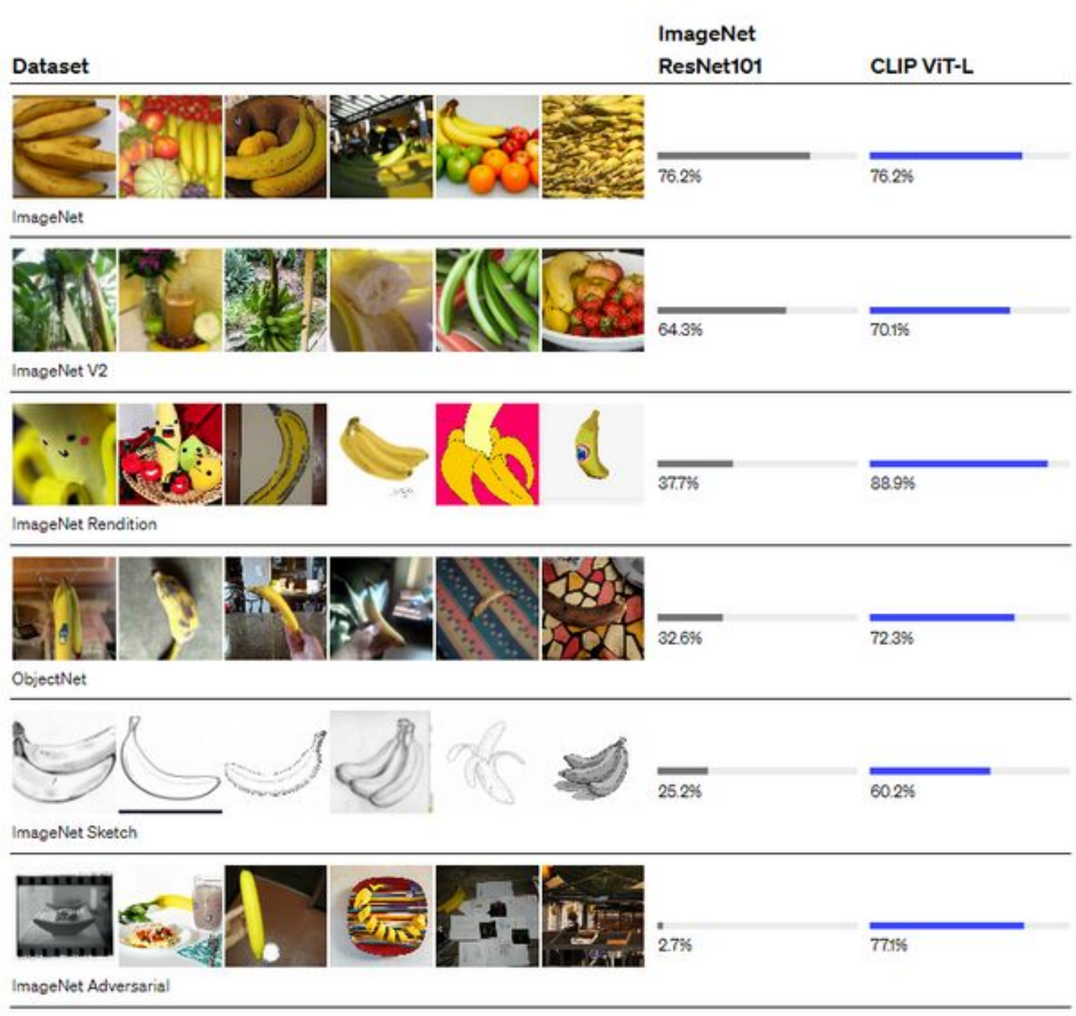
By
Risheendra
Mahika
Wisol
Высоцкий Иван Сергеевич

Why we need it

i)Harmful content
ii)Aesthetics

Why Clip

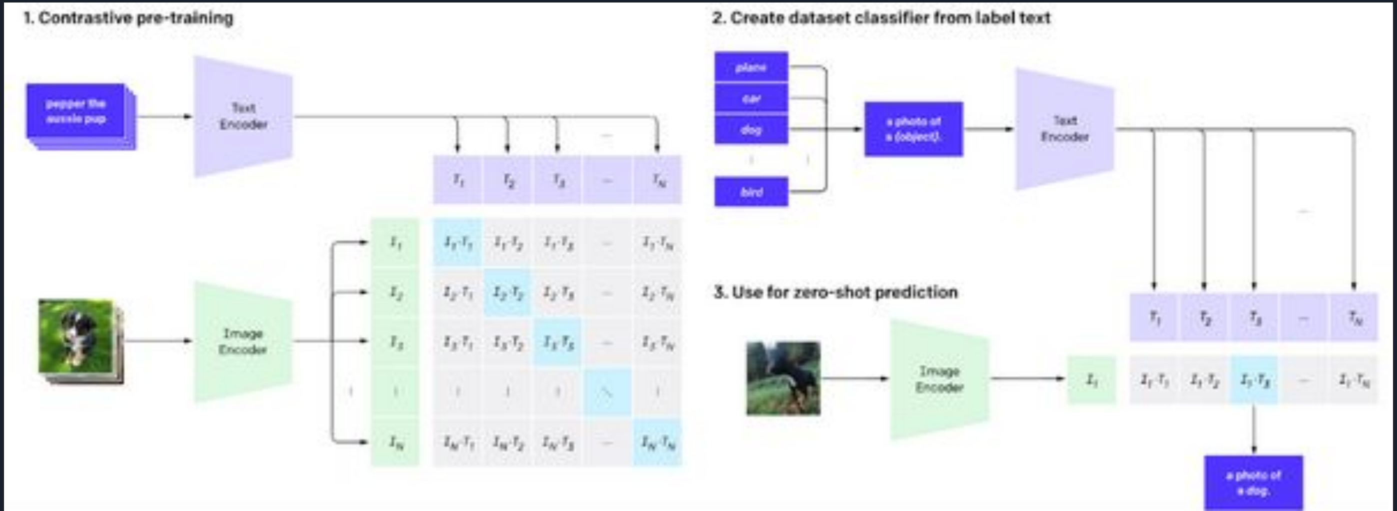
CLIP model is a versatile neural network that can assess image and video quality using a wide variety of images and natural language training.



Dataset	ImageNet ResNet101	CLIP ViT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

CLIP Creation

OpenAI has demonstrated that scaling a simple pre-training task enables the CLIP model to achieve competitive zero-shot performance on various image classification datasets. By utilizing text paired with images from the internet, CLIP learns to associate visual concepts with their names and can be applied to a wide range of visual classification tasks. For example, it can determine whether a given image is more likely to be paired with the text description "a photo of a dog" or "a photo of a cat" in datasets focused on classifying images of dogs vs. cats.

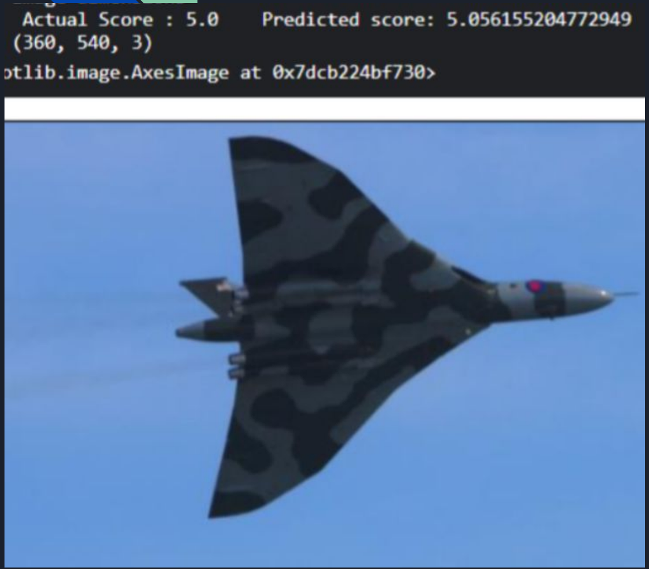

1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 … T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ … $I_1 \cdot T_N$
a photo of a dog.

	T_1	T_2	T_3	…	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	…	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	…	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	…	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	…	$I_N \cdot T_N$

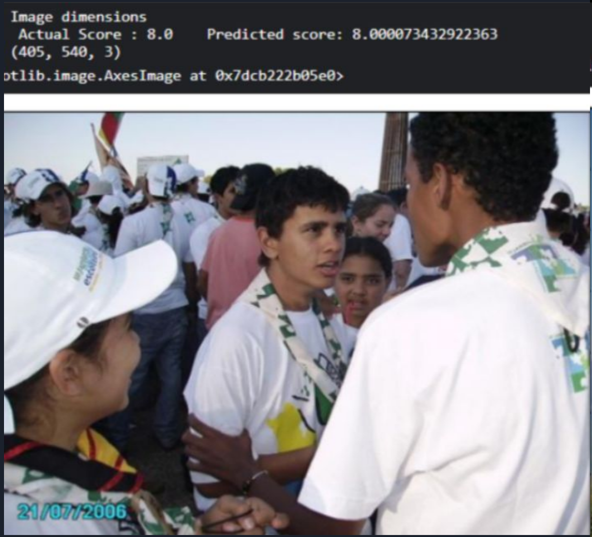
Advantages of clip

1. Filtering Images: CLIP can be used to filter a set of photos based on quality using a selected threshold.
2. CLIP can enhance image moderation by analyzing text and image inputs to detect and filter out inappropriate content, such as spam, violence, and explicit nudity, ensuring a safer digital environment
3. Regression Task: CLIP can predict image quality on a scale from 0 to 10, aligning with the assignment's requirements. The model's ability to consider both technical and aesthetic components of images makes it well-suited for this task

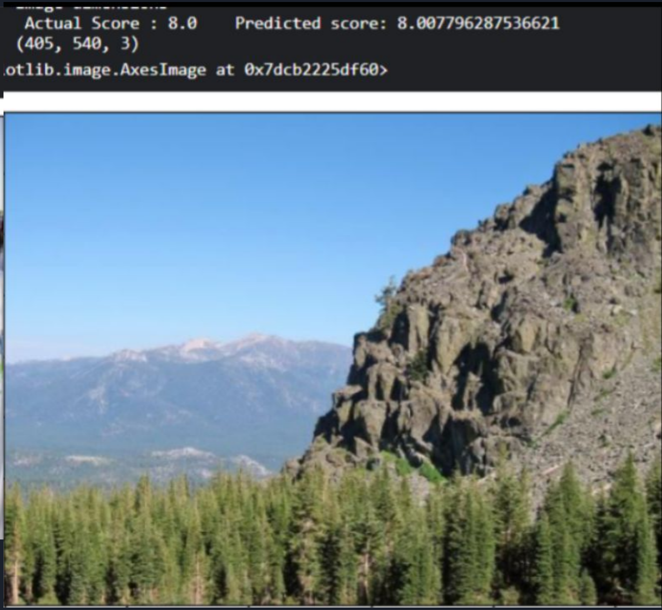
Examples



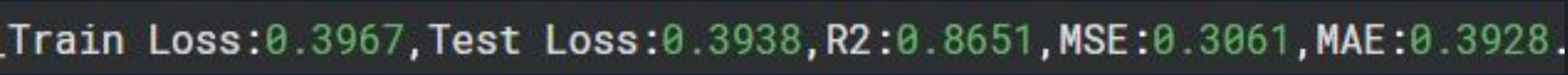
```
 Actual Score : 5.0    Predicted score: 5.056155204772949
(360, 540, 3)
otlib.image.AxesImage at 0x7dcb224bf730>
```



```
Image dimensions
 Actual Score : 8.0    Predicted score: 8.000073432922363
(405, 540, 3)
otlib.image.AxesImage at 0x7dcb222b05e0>
```



```
 Actual Score : 8.0    Predicted score: 8.007796287536621
(405, 540, 3)
otlib.image.AxesImage at 0x7dcb2225df60>
```



```
Train Loss:0.3967,Test Loss:0.3938,R2:0.8651,MSE:0.3061,MAE:0.3928.
```

Limitations

While CLIP excels at recognizing common objects, it struggles with more abstract or complex tasks such as counting objects or predicting proximity in images. It also faces difficulties in fine-grained classification tasks, such as distinguishing between car models or flower species. Moreover, CLIP exhibits poor generalization to images outside its pre-training dataset. For example, its performance on handwritten digits from the MNIST dataset falls short of human accuracy. Additionally, CLIP's zero-shot classifiers can be sensitive to wording and may require trial and error for optimal performance.

Broader impacts

CLIP empowers users to create their own classifiers, eliminating the need for task-specific training data. However, the design of these classes can strongly influence both model performance and biases. For example, when including terms like "criminal" or "animal" with Fairface race labels, CLIP tends to classify images of people aged 0-20 into the egregious category approximately 32.3% of the time. Interestingly, when adding the class "child" to the options, this behavior drops to around 8.7%.

Furthermore, CLIP's lack of reliance on task-specific training data enables it to handle niche tasks more easily. However, this also raises concerns regarding privacy and surveillance risks. For instance, CLIP achieves a top-1 accuracy of 59.2% for "in the wild" celebrity identification with 100 candidate choices, but this performance is not competitive compared to existing production-level models. OpenAI's paper delves deeper into the challenges posed by CLIP and highlights the importance of further research on understanding its capabilities, limitations, and biases.