

Learning Transferable Visual Models From Natural Language Supervision

<https://arxiv.org/pdf/2103.00020.pdf>

Under supervision of Dr. Jimson Mathew

By Mehuli Pal

mehuli_1901cs78@iitp.ac.in

OpenAI
ICML 2021

Contents

- Contrastive Learning
- Introduction to CLIP
- CLIP acts as a Bridge

Contrastive learning



Pig



Tiger



Panda



Hippo



Camel

Contrastive learning



Pig

Tiger

Panda

Hippo

Camel

AYAHOO

building (97.7%) Ranked 1 out of 12



✓ a photo of a **building**.

✗ a photo of a **carriage**.

✗ a photo of a **statue**.

✗ a photo of a **bag**.

✗ a photo of a **mug**.

IMAGENET BLURRY

marimba (79.5%) Ranked 1 out of 1000



✓ a photo of a **marimba**.

✗ a photo of a **abacus**.

✗ a photo of a **steel drum**.

✗ a photo of a **computer keyboard**.

✗ a photo of a **pool table**.

OBJECTNET IMAGENET OVERLAP

Pill bottle (98.3%) Ranked 1 out of 113



✓ a photo of a **pill bottle**.

✗ a photo of a **bottle cap**.

✗ a photo of a **beer bottle**.

✗ a photo of a **pillow**.

✗ a photo of a **wine bottle**.

DESCRIBABLE TEXTURES DATASET (DTD)

perforated (20.5%) Ranked 2 out of 47



✗ a photo of a **polka-dotted** texture.

✓ a photo of a **perforated** texture.

✗ a photo of a **dotted** texture.

✗ a photo of a **studded** texture.

✗ a photo of a **freckled** texture.

Introduction to CLIP

- Shorthand for **Contrastive Language-Image Pre-training**
- A neural network model built on hundreds of millions of images and captions
- Can return best caption given an image
- Has impressive **zero-shot** capabilities, making it able to accurately predict entire classes it's never seen before!

CLIP as a bridge between Computer Vision & Natural Language Processing

Thank You!..