CLIP’s performance on Bird Dataset

December 9, 2021

Motivation

Natural language descriptions and explanations have been used in NLP to create labeling functions to generate more labeled data (Hancock et al 2018, Snorkel)

For fine-grained classification tasks, like bird species classification, we often have plethora of language descriptions for the classes.

For bird species for example, many books and posts are dedicated to describing the species in detail, talking about its characteristics like feather color, beak, and etc.

Can we we combine this description with visual components to help generate image labels?

Previously Proposed Heuristic: using NLE to generate more labels for data

- Bird has red patch & black feather with white dots
- Red beak, grey feather, dot on feather

Example heuristic:

1. Create a grid of features & attributes that correspond to each class
2. **Calculate text similarity for the attributes to see which ones are most dissimilar (narrow down)**
3. Get semantic segment of the dissimilar features
4. **See how well the semantic segment matches up with the text attribute**
    a. (color match? Shape match?-- score)
    b. Option 1: image-text similarity
    c. Option 2: image -> text then text & text similarity
5. Give a label based on score
6. Choose 1 label among the labels

Previous Heuristic relies on image-text similarity score based on feature alignment.

How reliable would the image-text score be?

How well can we generate semantic segments based on text feature?

→ CLIP is a large-scale vision model trained by OpenAI that relies on “image-text” similarity

→ We look at CLIP to see how well it performs on birds dataset

Contrastive Language Image Pretraining (CLIP)

Creates a large (image, text) dataset called “WebImageText” & Trains Model with Contrastive objective

Dataset:

WebImageText: (400M text-image pairs from crawling the Web, 400x more than ImageNet)

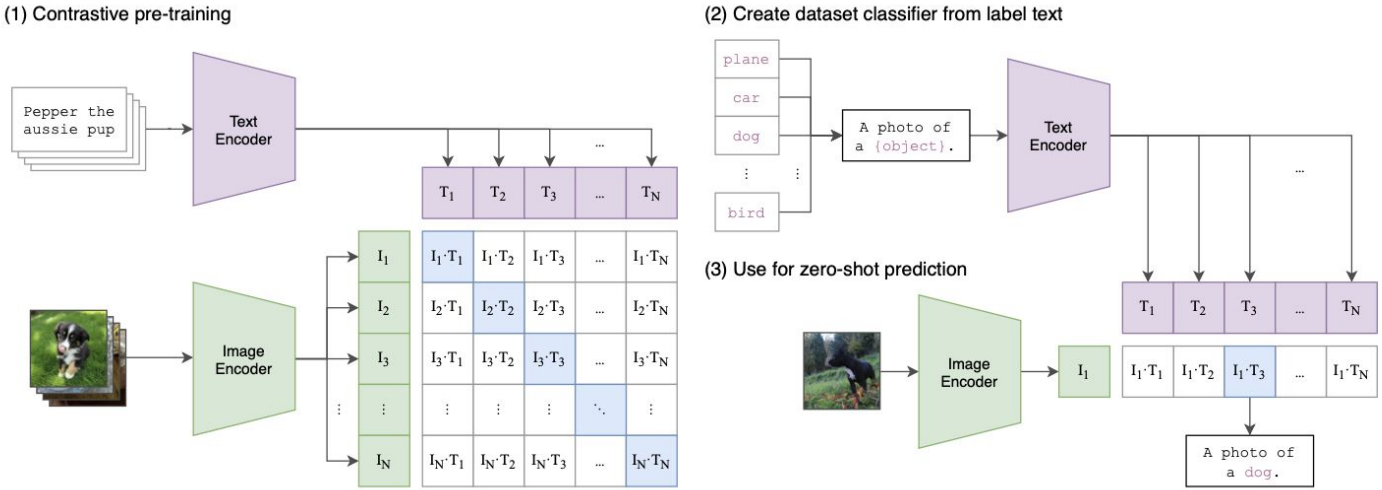
Pre-training method:

The two common approaches to pre-training visual models include image classification task and image captioning task

- But these are overly-reliant on the model’s ability to generate or label specific words or classes. As a result, they may end up omitting background information of an image, like the grass or clouds.
- CLIP pre-trains by providing (image, text) pairs and tries to distinguish which image belongs with which text as a whole, as opposed to a specific word.
    - *Provided motivation: It’s not forced to choose a restrictive label, so it retains background information better*

Model Architecture: Pre-Training Approach with contrastive objective

- N images, N texts → NxN grid of pairs
- Calculate Cosine Similarity of Image, Text Pairs
  - Maximize similarity of pairs (diagonal elements)
  - Minimize similarity of off-diagonal elements



Zero Shot Performance of CLIP

Pro: Matches performance of original ResNet-50 without being fed any of the 1.3 million ImageNet data.

Con: Not as good at complex tasks:

- Satellite image classification (EuroSAT, RESISC45)
- Lymph node tumor detection (PatchCamelyon)
- Counting objects in synthetic scenes (CLEVRCounts)
- German traffic sign recognition (GTSRB)
- Recognizing distance to the nearest car (KITTI Distance)

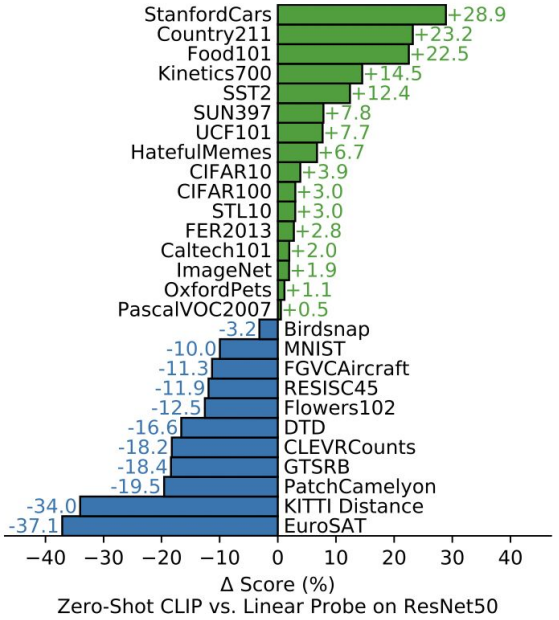


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

CLIP on Birds Dataset

Dataset:

CUB-200-211

15 semantic segments labeled, (Dtrain: 15 images per bird species)

Evaluation Metrics:

- Test Classification Accuracy (top 1, top 5)
- Test accuracy in hard to distinguish species
- Test accuracy in easy to distinguish species
- Test accuracy based on description used

Accuracy Result

CLIP is fed 5794 test images of 200 bird species

CLIP requires “text description” so the 200 class labels are coveted to 200 text descriptions. For each image, the text description with the highest score is the predicted class label for that image.

Overall, CLIP achieves accuracy of ~52%

Text Description	Top 1	Top 5
“A photo of a {class}”	52.9%	83.0%
“This is a photo of a {class}”	52.6%	82.9%

How well does CLIP distinguish between two similar species?

California Gull and Ring billed Gull are very similar.

In ALICE paper, contrastive natural language explanations from humans were used to aid in the visual classification of the two species.

A sample contrastive explanation was:

“Ring-billed gull has a bill with a black ring near the tip while California gull has a red spot near the tip of lower mandible”

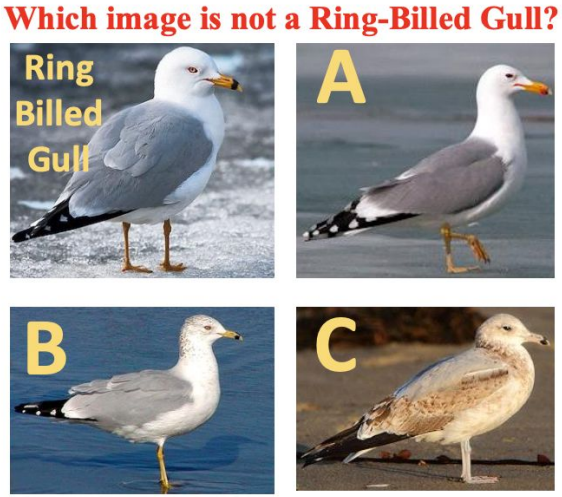


Figure 1: An example task that would benefit from learning with natural language explanation. The top-left corner shows an example image of a ring-billed gull. In the other three images (A), (B), (C), which one is not a ring-billed gull but a California gull? Given the natural language explanation *“Ring-billed gull has a bill with a black ring near the tip while California gull has a red spot near the tip of lower mandible”*, it would be easier to find that (A) is the correct choice.

CLIP is also reliant on class name information

If the text description simply uses the class names, we get 70% accuracy with just class name as the description. (120 images, 84 identified accurately)

	Text Description	Accuracy
California gull	"this is a california gull"	70%
Ring billed gull	"this is a ring billed gull"	

If we use the human provided contrastive explanation, CLIP achieves 50-51% accuracy.

	Text Description	Accuracy
California gull	"this is a bird with red spot near the tip of its lower mandible"	50-51%
Ring billed gull	"this is a bird with with black ring near the tip of its bill"	

Using simplified description did not improve CLIP performance

If we use the human provided contrastive explanation, CLIP achieves 50-51% accuracy.

	Text Description	Accuracy
California gull	"this is a bird with red spot near the tip of its lower mandible"	50-51%
Ring billed gull	"this is a bird with with black ring near the tip of its bill"	

If the text description uses simplified versions of contrastive explanation, we get 48.3-49.1% accuracy

	Text Description	Accuracy
California gull	"this is a bird with red spot near its bill"	48- 49%
Ring billed gull	"this is a bird with with black ring near its bill"	

Can CLIP distinguish between red and grey neck?

The contrastive feature for california gull and ring-billed gull may be too complex.

How does CLIP perform for simpler distinction?

Black footed albatross bird and Laysan Albatross bird are very similar except foR their neck. One has grey neck and another has white neck.

laysan_albatross
a photo of laysan albatross bird

black_footed_albatross
a photo of black footed albatross bird

Text description leads to lower accuracy

If the text description simply uses the class names, we get 93% accuracy with just class name as the description. (120 images, 112 identified accurately)

	Text Description	Accuracy
Black footed albatross	"this is a black footed albatross bird”	93%
Laysan Albatross bird	"this is a laysan albatross bird”	

If we use feature based description, CLIP achieves 75% accuracy.

	Text Description	Accuracy
Black footed albatross	this is a bird with grey neck	75%
Laysan Albatross bird	"this is a bird with white neck"	

It’s possible that image variation partly contributes to lower accuracy and that CLIP has already seen these images before

In this image, you can’t really see the bird’s body.

Regardless, CLIP seems to be able to distinguish between the two species.

CLIP data comes from crawling the web.

It’s possible that CLIP has already seen these images of birds before when it was being trained?