CLIP Evaluation

Rabia Aslam , Abdul Samad , Muhammad Salman

***Abstract*—We have evaluated the performance of Contrastive Model CLIP with (ViT-B/32) which was pretrained on 400M image-text pairs across multiple datasets including CIFAR-10, STL-10 and PACS. We have focused on zero-shot accuracy of CLIP's classification. We experimented with prompt engineering, comparing single-word class labels with natural language text descriptions. Results show that rich prompts produce better accuracy. To gauge robustness of CLIP on domain-shifted data, we tested out-of-distribution datasets from the Model-VS-Human benchmark. Our findings show semantic biases in CLIP's results and confirm that CLIP has strong shape bias as it gives high accuracy on edge datasets versus cue-conflict. We also created custom datasets to evaluate preference of CLIP on shape vs. color and texture vs. shape biases. A low score on cue-conflict may be attributed to dataset complexity, but CLIP outperformed CNNs in correctly classifying objects. Lastly, we tested text retrieval based on image and image retrieval based on text.**

I. Introduction

Recent research is focused on domain generalization. With our experiments, we evaluated the robustness of CLIP on OOD and manipulated datasets. We tested CLIP on the Model-vs-Human benchmark. The datasets evaluated include edge/shape, contrast, cue-conflict and silhouette. CLIP achieved above 80% accuracy on shapes data, which is impressive given that CLIP was never trained on this data.

Research Questions

1) Zero-shot performance of CLIP on diverse datasets.
2) Role of prompt engineering on zero-shot accuracy.
3) Semantic biases in CLIP vs. CNNs.
4) Robustness of CLIP on out-of-distribution data.
5) Robustness of CLIP on highly manipulated data.

II. Experiments

A. Zero-shot classification on CIFAR-10, STL 10 and PACS

To evaluate zero-shot accuracy of CLIP on these datasets we have calculated the cosine similarity between image embeddings and class text embedding for CIFAR-10, STL-10 and PACS dataset and have compared these accuracies with that of ResNet.

B. Effect of Prompt Engineering

Rich prompts produced better accuracy than single-word prompts. We defined multiple prompt templates and evaluated per-template and ensemble prompt accuracies.

C. Robustness on OOD Data

We used the Model-Vs-Human benchmark, which probes model biases. Datasets included:

- Edge/Shape (shape-based classification)
- Cue-Conflict (texture vs. shape conflict)
- Silhouette (global shape without texture)

D. Custom Bias Probing

We have created some custom datasets to understand and evaluate the behavior of CLIP . we have taken some images with varying colors,shapes, some drawings, sketches and photos and have projected them with t-SNE to show that Clip understands semantics of image and align images of same class together irrespective of their color or domain (drawing, photo, sketch).

E. Cross-Modal Retrieval

We have performed some custom experiments for better visualizations of how CLIP encodes texts and images and how it calculates cosine similarity when the dataset is very limited. We have taken 23 images and 36 corresponding texts. And have CLIP compute the similarity of each image with text. Afterwards we have done Image to Text and Text to Image retrieval.

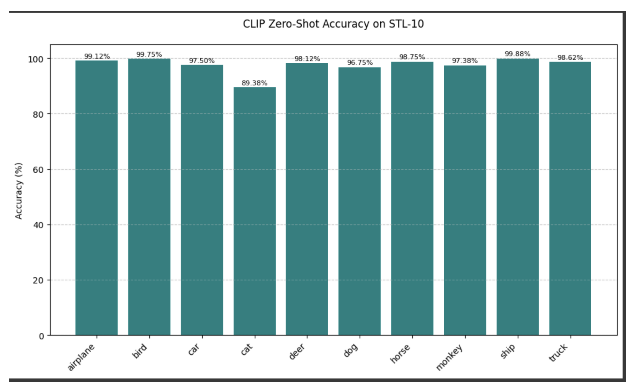

Fig. 1: Zero-shot results on STL-10 dataset.

III. Results and Discussion

A. CLIP and Contrastive Biases

CLIP achieved $> 80\%$ accuracy on CIFAR-10, STL-10, and PACS benchmarks. CLIP was tested on 1000 test images of CIFAR and 8000 test images of STL. Although CLIP was never explicitly trained on these datasets, the high accuracy on CIFAR-10 and STL-10 can be attributed to the common class range. As CLIP has seen 400M image-text pairs during its training, it has mostly correctly classified CIFAR images..Fig 1

B. Prompt Engineering

We evaluated CIFAR-10 accuracy on one word class names vs ensemble prompts which were custom defined in code and have seen that ensemble prompts increase accuracy. This is the same as different people telling about one object increases our understanding of the concept so has happened in case on CLIP.Fig 2

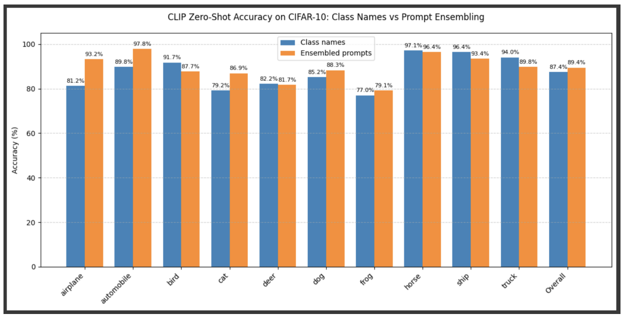

Fig. 2: Effect of prompt engineering on CIFAR-10 zero-shot accuracy.

C. Comparison of Accuracies across models

The chart shows that trained ViT-S/16 has outperformed CLIP (ViT-B/32) . so training increases accuracy on the dataset it is trained on.Fig **??**

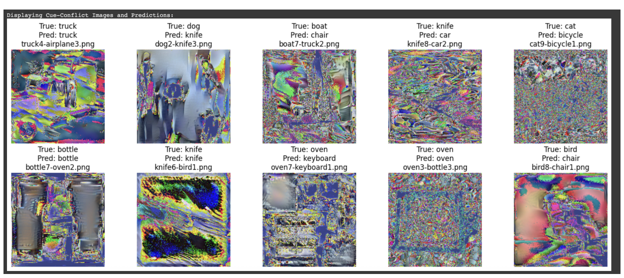

Fig. 3: CLIP performance on cue-conflict dataset (example 2).

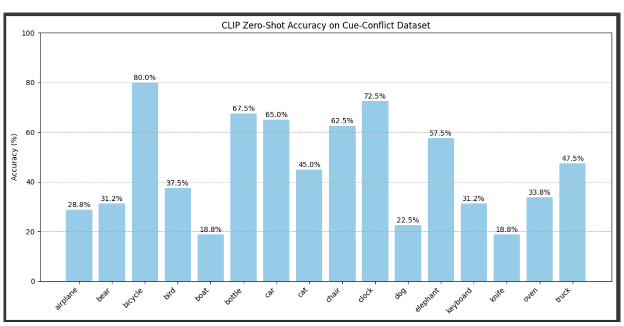

Fig. 4: Zero-shot cue-conflict results.

D. Evaluation on PACS Dataset

The PACS dataset contains images of four domains i.e. Photos, Drawings, Sketches and Paintings. Evaluating CLIP on PACS defines domain generalization as to how accurately CLIP can handle domain shift. On Photos we get accuracy of 99.82 which should be so as clip was mainly trained on internet images and a large number of images on internet are photos. It performed poorest (though still a very good accuracy just poor w.r.t other domains) with accuracy of 86which can be attributed to the fact that sketches lacks color and texture and although clip is shape biased but sketches can be very diverse and show different or conflicting concepts. Although CLIP is shape biased but a relative low accuracy on sketches suggests that CLIP does check color and texture along with Shapes to get excellent semantics of images. A high accuracy on paintings is because they are nearest to pictures while drawing basically extracts shapes and CLIP is known to be shape biased. Hence proved as well.

Below is a diagram of dataset images that can be found in PACS dataset and the charts showing accuracy of CLIP in four domains. Fig5,6

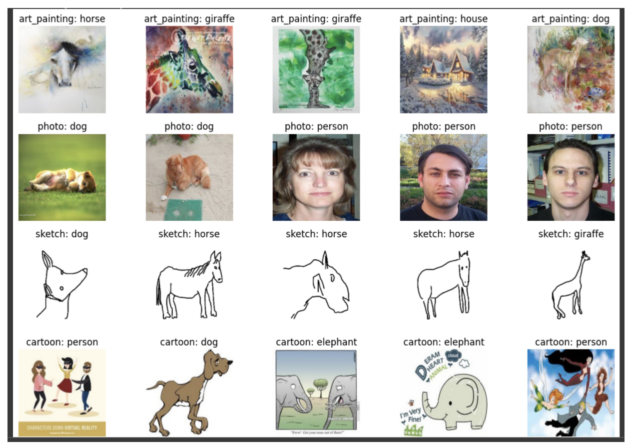

Fig. 5: PACS dataset domains.

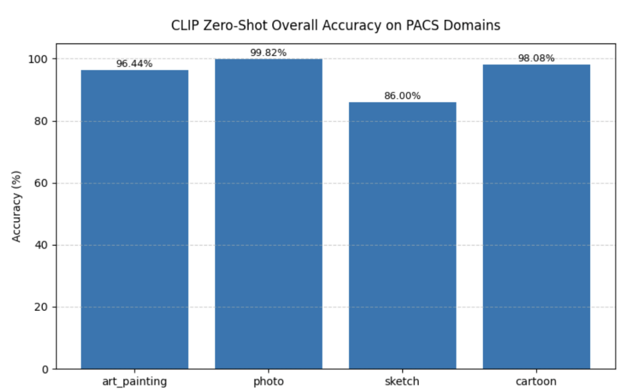

Fig. 6: Zero-shot PACS domain results.

We have also evaluated accuracy of CLIP on class level of different domains and has observed that clip had difficulty in correctly classifying class of “Dog” and “Giraffe” the remaining high accuracies on different domains strengthen domain generalization of CLIP.Fig7

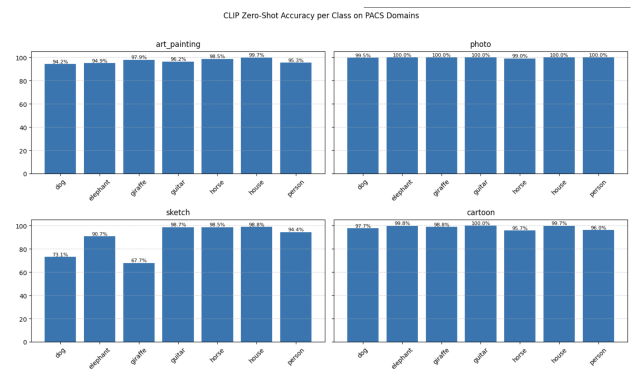

Fig. 7: Domain-level zero-shot PACS accuracy.

IV. Evaluation of Inductive Biases in CLIP

A. Evaluation of CLIP on CUE-CONFLICT data:

Evaluation of CLIP on CUE-CONFLICT data: The cue-conflict data is taken from Model-vs-Human benchmark to evaluate Shape vs Texture bias in CLIP . CLIP has shown poor performance mostly below 60and worst for the class “BOAT” . But to be honest this dataset seem quite hard to guess, as a human i m sure my own accuracy will be lower than CLIP .(Figs. **??**–3– 4).

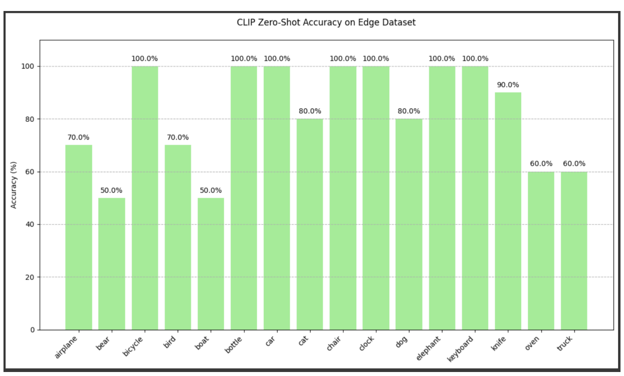

Fig. 8: Zero-shot accuracy on edge dataset.

B. Evaluation Of CLIP on edge images for Shape Bias

In 7 out of 10 cases Clip has given 100% accuracy and mostly accuracy is in range 70 to 90. With only 50 for bear and boat. This proves CLIP to be shape biased. CLIP seems to be superconfused about boat and beer where it mostly classifies boat as knife as without color and texture and with low resolution of images this is an expected result.Fig 9,8

C. Evaluation Of CLIP on Silhouettes:

CLIP has shown diverse range of accuracies on silhouettes ranging from 10 to 100. This is because silhouettes just gives boundaries with no edges and distinguishing cues. The dataset

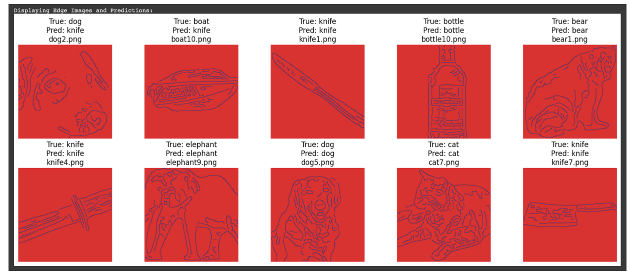

Fig. 9: CLIP performance on edge dataset (example images).

images as can be seen are highly diverse, cropped, rotated and noisy. But still CLIP has performed well for 9 out of 16 classes. Which again strengthens the shape bias of CLIP . has CLIP been texture biased we would have seen very low accuracies.(Figs. 10–11).

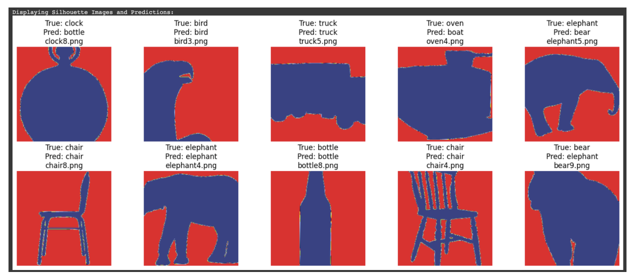

Fig. 10: CLIP results on silhouette dataset (example images).

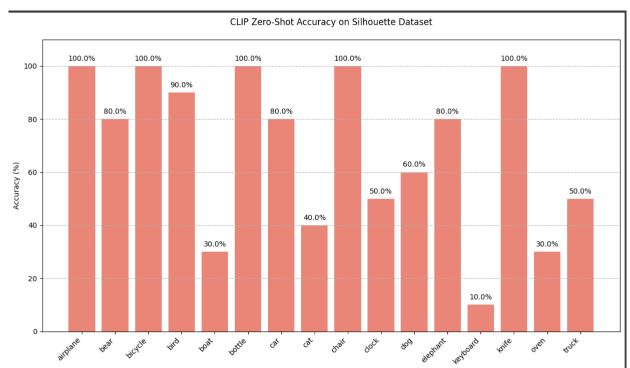

Fig. 11: Zero-shot accuracy on silhouettes dataset.

D. CLIP Image to text retrieval:

We have downloaded some images (23 in number as are uploaded on Github as 23 images data) from the internet and have written some diverse text statements (texts for test.csv) to evaluate CLIP on our custom dataset. The aim was to see that if we have a very small dataset and very few relevant text image pairs then how CLIP performs. The results were extremely good and can be seen from the image. As an example for the image of four cats sitting in a row, CLIP has extracted texts “A cute kitten” , “A happy group of people”. We do had other text encodings like “cute kitten in hand” and “baby kitten playing” but it seems CLIP recognized that there was no hand in image and the cats were bigger than kittens and so as the second text it retrieved " A happy group of people". Often on the internet people do upload such metaphorical statements so its understandable and admirable that CLIP recognizes the fact and retrieves the best possible match. Similarly in the dataset we had people working in the office, people working and kitchen crew. For the image of kitchen crew , CLIP has correctly retrieved texts and has not confused kitchen with office which shows that CLIP understands backgrounds of images.Fig 12

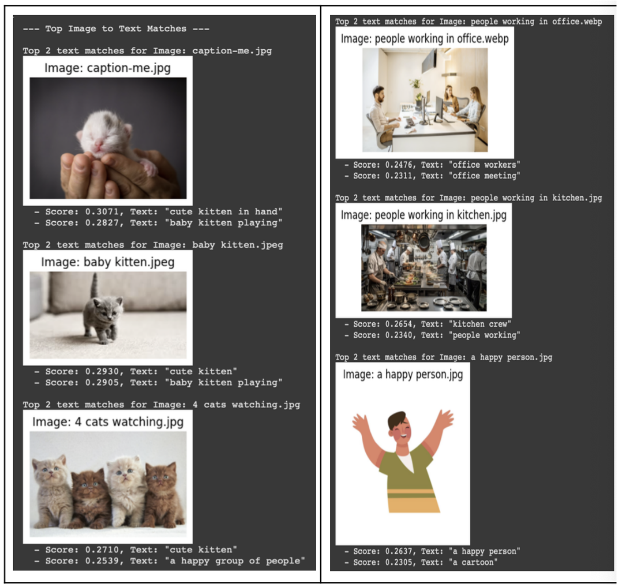

Fig. 12: CLIP image-to-text retrieval examples.

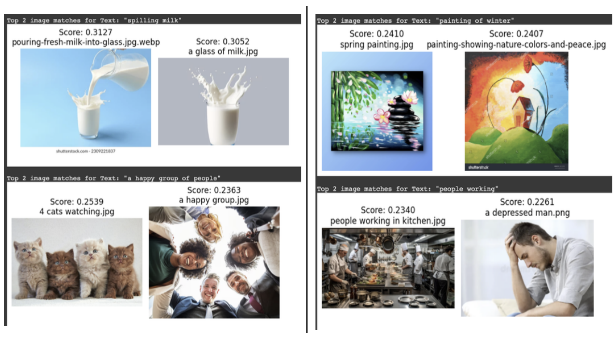

Fig. 13: CLIP text-to-image retrieval examples.

Next we experimented with image retrieval based on texts from our same custom dataset and it is acknowledgeable that despite the fact that images of water glasses were also part of dataset, CLIP has correctly retrieved images of milk glasses. This suggests that although CLIP is widely known to be shape biased it does sees texture and colors and correctly identifies it. We did this experiment 2 waysm, since we had only 3 images of milk in dataset so if for the text prompt we retrieve four images and last image will be of water class. This again shows that CLIP understood the semantics and brought the closest pair out.Fig13

E. T-SNE Projection of CLIP Image Embeddings

Our representational analysis highlights both the strengths and limitations of CLIP's embedding space when applied to domain adaptation. We plotted t-SNE projection of our custom dataset which is included in our github repo under "representational analysis". The projection shows domain adaptation of CLIP. despite being part of different domains (photos, sketch, drawing, Silhouettes) images of the same class are close in embedding space. Although we do see some misplacements as well. For similarity complexity , we have added silhouettes of goat, zebra, horse, donkey and deer and it's evident that horse, zebra and donkey are closer in embedding space despite being part of different classes. Which shows that mere silhouettes doesn't give enough visual cues to differentiate among classes which have similar shapes. We also experimented with shapes of different colours and proved that CLIP is shape biased as images of circles are clustered together despite the color, and images of rectangles are away from triangles and circles. the findings imply that while CLIP effectively captures cross-domain semantic similarity, its reliance on shape over other visual attributes can lead to confusions in tasks where classes share similar silhouettes or contours.Fig14

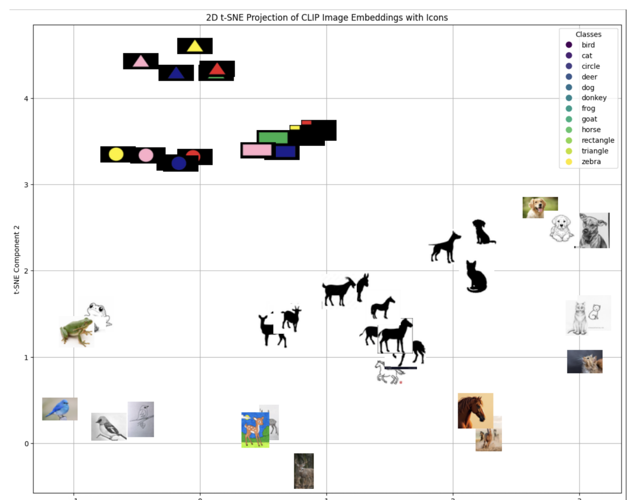

Fig. 14: CLIP text-to-image retrieval examples.

V. Conclusion

Our experiments show that CLIP is strongly shape-biased, while CNNs tend toward texture bias. Trained models outperform CLIP on in-domain datasets, but CLIP demonstrates impressive domain generalization, confirming its robustness and multimodal capabilities.