Zero Shot Image Classification On VizWiz Image Classification Dataset

Basim Baqai[a], **Maisum Abbas**[a] **and Abdul Rehman Nazeer**[a]

[a]*Department of Artificial Intelligence*

Bilal Ahsan - Spring 2025

Abstract—This study explores zero-shot image classification performance on the VizWiz Image Classification Dataset, which features photos taken by visually impaired individuals. We recreate and extend the work from Bafghi et al. (CVPR 2023), comparing the performance of advanced neural network architectures (AlexNet, VGG11, ViT, and CLIP) across multiple datasets: ImageNet, ImageNetV2, ObjectNet, and VizWiz. Our analysis focuses particularly on CLIP's zero-shot capabilities and its applicability to accessibility-focused datasets. Through prompt engineering and embedding-based retrieval methods, we demonstrate how these models perform on images captured in challenging, real-world conditions by blind photographers, highlighting both the strengths and limitations of current zero-shot classification approaches.

Keywords—*Zero-shot learning, image classification, CLIP, VizWiz dataset, computer vision, embeddings*

1. Introduction

Visual recognition systems have made remarkable progress in recent years, but their performance on diverse, real-world data remains challenging, particularly for images captured by individuals with visual impairments. The VizWiz dataset [8] presents unique difficulties due to its unconventional image characteristics - including blur, poor framing, and unusual angles - that are not typically seen in standard computer vision datasets.

In this study, we investigate zero-shot image classification approaches on the VizWiz Image Classification Dataset, which contains images taken by blind photographers. Zero-shot learning is particularly valuable in this context as it eliminates the need for extensive labeled data from specialized populations. We evaluated several state-of-the-art models, including AlexNet [3], VGG11 [5], Vision Transformer (ViT) [6], and CLIP [7], comparing their performance with multiple benchmark datasets.

2. Related Work

2.1. VizWiz Dataset

The VizWiz Image Classification Dataset was introduced by Bafghi et al. [8] as a benchmark to evaluate image classification models on pictures taken by blind individuals. Unlike curated datasets like ImageNet [1], VizWiz reflects the authentic challenges faced by visually impaired users, including motion blur, poor lighting, and partial object visibility. This dataset has become an important benchmark for evaluating visual recognition systems in the context of accessibility.

2.2. Zero-Shot Image Classification

Zero-shot learning has emerged as a promising direction for image classification without requiring examples of all possible classes during training. Early approaches relied on attribute-based descriptions [2], while more recent methods leverage pre-trained language models to establish connections between visual inputs and textual labels [4]. CLIP [7] represents a significant advancement in this area, using contrastive learning between image and text embeddings to achieve impressive zero-shot performance across various visual tasks.

2.3. Vision Models

We evaluate several key vision architectures in our study: AlexNet [3], which popularized deep convolutional networks for image classification; VGG11 [5], known for its uniform architecture of stacked convolutional layers; Vision Transformer (ViT) [6], which adapts transformer architectures from NLP to computer vision tasks; and CLIP [7], which jointly trains image and text encoders on large-scale web data.

3. Methodology

3.1. Datasets

Our experiments utilize four datasets:

- **ImageNet**: The standard benchmark with 1,000 object categories.
- **ImageNet V2**: A new test set following ImageNet's protocol but with different images.
- **ObjectNet**: A test set focusing on objects in challenging poses and contexts.
- **VizWiz**: Images captured by blind individuals, presenting unique accessibility challenges.

3.2. Model Architecture Comparison

We evaluated four deep learning architectures:

- **AlexNet**: An early CNN architecture with 5 convolutional layers.
- **VGG11**: A deeper network with uniform 3x3 convolutional filters.
- **ViT**: A transformer-based approach that processes image patches.
- **CLIP**: A dual-encoder model combining visual and textual information.

For all models except CLIP, we utilized pre-trained weights from ImageNet and evaluated their transfer learning capabilities to other datasets. For CLIP, we leveraged its zero-shot classification abilities through prompt engineering.

3.3. CLIP-based Zero-Shot Classification

Our CLIP implementation follows these key steps:

3.3.1. Prompt Engineering

We formatted class labels to match CLIP's pre-training by creating simple prompts: "a photo of a [class]" for each ImageNet class.

3.3.2. Embedding Creation

For both text prompts and images:

- Text prompts were tokenized and encoded using CLIP's text encoder.
- Test images were processed and encoded using CLIP's image encoder.
- All embeddings were normalized to unit length.

3.3.3. K-Nearest Neighbors Voting

We implemented a k-NN voting approach:

- For each test image, we found the k most similar training images.
- Each similar image "voted" for a class based on its embedding's similarity to text prompts.
- The class with the most votes became the final prediction.

4. Implementation

Our implementation leverages PyTorch for model evaluation. In the following, we highlight key components of our CLIP-based classification pipeline.

4.1. Prompt Engineering and Text Embedding

We generate prompts by adding the prefix "a photo of a" to each ImageNet class label:

```
prompts = [f"a photo of a {name}" for name in imagenet_classes]
with torch.no_grad():
    text_tokens = clip.tokenize(prompts).to(device)
    text_features = model.encode_text(text_tokens)
    text_features /= text_features.norm(dim=-1, keepdim=True)
```

4.2. Image Embedding and Retrieval

For test images, we compute embeddings and compare them against a database of precomputed training embeddings:

```
with open(train_embedding_path, "rb") as f:
    train_embeddings_dict = pickle.load(f)
    filenames = list(train_embeddings_dict.keys())
    train_embeddings = torch.stack([train_embeddings_dict[f] for f in filenames]).to(device)
    train_embeddings /= train_embeddings.norm(dim=-1, keepdim=True)

test_image = preprocess(Image.open(test_image_path).convert("RGB")).unsqueeze(0).to(device)
with torch.no_grad():
    test_feature = model.encode_image(test_image)
    test_feature /= test_feature.norm(dim=-1, keepdim=True)
```

4.3. K-Nearest Neighbors Voting

We identify the k most similar training images and implement a voting scheme:

```
similarities = test_feature @ train_embeddings.T
topk_sim, topk_idx = similarities.topk(k, dim=1)
Implement voting
votes = []
for idx in topk_idx[0]:
    img_feat = train_embeddings[idx].unsqueeze(0)
    text_sim = img_feat @ text_features.T
    best_text_idx = text_sim.argmax().item()
    votes.append(imagenet_classes[best_text_idx])

vote_counts = Counter(votes)
predicted_class, count = vote_counts.most_common(1)[0]
```

5. Limitations

While our work provides valuable insights into using CLIP for image classification in assistive contexts, several limitations remain:

- **Task Scope:** We focus exclusively on image classification. However, real-world assistive technologies often require more complex tasks such as object detection or instance segmentation to provide meaningful feedback to users.
- **Computational Considerations:** Our analysis does not take into account computational efficiency, which is critical for deployment on mobile or embedded devices commonly used by visually impaired individuals.
- **Dataset Constraints:** Although we employ diverse datasets, they still capture only a limited subset of real-world environments and scenarios encountered by users with visual impairments.
- **Domain Misalignment:** The pretrained CLIP model exhibits poor domain alignment with assistive tasks, as it was trained on generic web data that may not reflect the visual content relevant to assistive use cases.
- **Weak User Context Understanding:** The system lacks mechanisms to model or infer user intent or context, which limits its usefulness in practical, dynamic assistive settings.
- **Voting Bias:** Our k-NN voting mechanism may suffer from class imbalance and bias in the training data, leading to skewed predictions in ambiguous or unseen scenarios.

6. Experimental Results Analysis

- **Standard ImageNet:** All four vision models—AlexNet, VGG11, ViT, and CLIP—performed well on standard ImageNet, demonstrating high classification accuracy and serving as a baseline to assess their competence under ideal and well-curated conditions.
- **ImageNetV2 (Distribution Shift):** When evaluated on ImageNetV2, which introduces slight distributional shifts, AlexNet exhibited a noticeable drop in performance. In contrast, VGG11, ViT, and CLIP maintained strong accuracy, showing better generalization across similar but shifted data.
- **ObjectNet (Viewpoint and Context Variation):** ObjectNet challenged the spatial and contextual robustness of the models. The convolutional networks (AlexNet and VGG11) produced several incorrect predictions due to changes in object orientation and background context. Transformer-based models (ViT and CLIP), however, remained robust and showed significantly higher classification accuracy under these variations.
- **VizWiz (Accessibility-Focused Dataset):** The VizWiz dataset, composed of low-quality and often poorly framed images taken by blind users, proved the most difficult. All models produced divergent and frequently incorrect predictions for the same images. This inconsistency highlights the severe challenges current vision architectures face in accessibility-related, real-world applications.
  - **Standard CLIP:** 62% accuracy.
  - **Voting-based CLIP:** 74% accuracy.

7. Conclusion

We evaluated several deep learning models for zero-shot image classification on the VizWiz dataset, which contains photos taken by blind users. Among them, CLIP demonstrated the highest robustness, highlighting the benefits of large-scale multimodal pretraining for generalization to real-world assistive scenarios.

Despite this, all models showed significant performance drops compared to standard datasets, emphasizing the need for vision systems trained and tested on data that reflects the experiences of visually impaired users. This gap reinforces the importance of domain-specific evaluation.

Future work should focus on adapting models to the unique visual patterns in assistive datasets, incorporating user context, and extending beyond classification to object detection and segmentation. These steps are critical to building reliable and practical assistive technologies.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", pp. 248–255, 2009. DOI: 10.1109/CVPR.2009.5206848.

[2] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by transferring discriminative object parts", pp. 2168–2175, 2009. DOI: 10.1109/CVPR.2009.5206694.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, USA*, 2012, pp. 1097–1105. [Online]. Available: https : / / proceedings . neurips . cc / paper / 2012 / file / c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[4] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer", pp. 2035–2043, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/file/f6b2577a71dbd52065ca0efcd7a3d788-Paper.pdf.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.

[6] A. Dosovitskiy, L. Beyer, N. Houlsby, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[7] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision", *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020.

[8] Y. H. Bafghi, S. Mohseni, M. H. Rohban, A. Jahanian, R. Zemel, and Y. J. Lee, "Vizwiz-classification: Boosting accessibility for blind users via image classification", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023, pp. 19 866–19 876. DOI: 10.1109/CVPR52729.2023.01922. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Bafghi_VizWiz-Classification_Boosting_Accessibility_for_Blind_Users_via_Image_Classification_CVPR_2023_paper.html.