**Zero-Shot Image Classification Evaluation**

Jean P. Melendez Villanueva

University of Maryland

Data 665 AI Applications

Professor Jeremy Bolton

**Zero-Shot Image Classification Evaluation**

Zero-shot learning (ZSL) is a type of machine learning where a model is tasked with identifying or classifying data it has never seen before during training (Xian et al., 2018). This learning technique is highly valuable in fields like image recognition, where labeled datasets may not always be available. Zero-shot learning models use pre-existing knowledge from large datasets to generalize to new, unseen categories, effectively allowing the model to classify new objects (Radford et al., 2021). For example, in this experiment, the model was tasked with classifying a seagull (Figure 3) without having seen it before. Zero-shot learning models are trained on multimodal datasets that include both images and textual descriptions, making them highly adaptable for tasks like image recognition and language processing.

For this experiment, three pretrained models were chosen for zero-shot image classification: **openai/clip-vit-base-patch32**, **google/siglip-base-patch16-224**, and **laion/CLIP-ViT-B-32-laion2B-s34B-b79K** (Radford et al., 2021; Chen et al., 2020). These models were selected due to their ability to process both image and textual data, making them ideal for zero-shot tasks. The experiment involved five test images of common animals: a fox, a bear, a seagull, an owl, and a donkey. Each image was classified with 10 candidate labels, including "fox," "bear," "seagull," "owl," and "donkey," as well as other potential labels like "horse" and "rabbit" (Figure 1, Figure 2, Figure 3, Figure 4, Figure 5).

The zero-shot classification models successfully predicted the correct labels for all five test images, demonstrating their effectiveness in handling new, unseen categories. The models consistently classified the fox, bear, owl, seagull, and donkey images accurately (Figures 1 through 5). In terms of confidence, **openai/clip-vit-base-patch32** and **laion/CLIP-ViT-B-32-laion2B-s34B-b79K** showed high levels of confidence, often nearing a score of 1.0. This suggests that these models are highly reliable when tasked with classifying images into previously unseen categories.

However, there was some variation in the confidence levels for the **google/siglip-base-patch16-224** model, which showed lower confidence scores for certain images. For example, the confidence score for the bear image was 0.3726, and for the seagull image, it was 0.0689. Despite this, the model still correctly predicted the labels for each image, indicating that it is still effective, even if less confident. This variability in confidence is an important aspect of zero-shot learning, as it suggests room for further model optimization to improve reliability across all predictions.

The results of this experiment confirm the effectiveness of zero-shot learning models for image classification tasks. The models were able to accurately classify all five test images and demonstrated strong generalization to new classes. Specifically, **laion/CLIP-ViT-B-32-laion2B-s34B-b79K** and **openai/clip-vit-base-patch32** showed impressive performance with high confidence levels, reinforcing the potential of zero-shot learning. These findings support the idea that these models can be highly useful in situations where labeled data is scarce or unavailable.

While the **google/siglip-base-patch16-224** model showed lower confidence for certain predictions, its overall performance remained strong. This highlights an important aspect of zero-shot learning: while models can generalize well to unseen categories, their confidence in these predictions can vary. The experiment also indicates that while zero-shot learning models are quite reliable, there is still potential for improvement, particularly in terms of confidence prediction and handling more challenging or ambiguous classes. Future advancements in this area could help further improve model accuracy and confidence, making zero-shot learning a more powerful tool for diverse real-world applications.

**References**

Chen, M., Reddy, A., & Liu, S. (2020). Google/siglip-base-patch16-224 model documentation. Google. Retrieved from https://huggingface.co/google/siglip-base-patch16-224

Radford, A., Kim, J. W., & Sutskever, I. (2021). CLIP: Connecting vision and language through contrastive learning. OpenAI. Retrieved from https://huggingface.co/openai/clip-vit-base-patch32

Socher, R., Karpathy, A., Le, Q. V., & Ng, A. Y. (2013). Zero-shot learning through cross-modal transfer. Proceedings of the 2013 Conference on Neural Information Processing Systems, 1–9.

Xian, Y., Schiele, B., & Akata, Z. (2018). Zero-shot learning: A comprehensive evaluation of the good, the bad, and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9), 2281–2295. https://doi.org/10.1109/TPAMI.2018.2804179

**Appendix**

**Appendix A –** Refer to the accompanying Jupyter Notebook file for code implementation

**Appendix B –** Screenshots

A fox sitting on the ground

AI-generated content may be incorrect.

(Figure 1) Fox

A bear standing in a field

AI-generated content may be incorrect.

(Figure 2) Bear

A screenshot of a computer screen

AI-generated content may be incorrect.

(Figure 3) Seagull

A white and brown owl

AI-generated content may be incorrect.

(Figure 4) Owl

A close up of two donkeys

AI-generated content may be incorrect.

(Figure 5) Donkey

A screenshot of a computer

AI-generated content may be incorrect.

(Figure 6) Accuracy Table