Evaluating and Improving Calibration in Zero-Shot Text-Image Classification Models

Ayush Agarwal
aagarw41@jhu.edu

Kahnrad Braxton
kbraxto6@jhu.edu

Cameron Carpenter
ccarpe18@jhu.edu

William Shiber
wshiber1@jhu.edu

Abstract

Zero-shot image classifiers, such as those built using CLIP and SigLIP, allow image classification without requiring labeled training data. These models compare feature vectors derived from images and text, but their raw outputs often fail to provide reliable probability estimates. In this study, we explore methods like Platt Scaling, Isotonic Regression, and Similarity Binning Averaging to improve the interpretability and trustworthiness of their predictions.

A key focus of our work is understanding whether calibration parameters can transfer across different domains and classes. For instance, we investigate whether a calibrator trained for identifying bridges can also be applied to other tasks like identifying airfields while still providing trustworthy probability estimates. We test these methods across various conditions, including cross-class and cross-domain image analysis.

Our contributions include the construction of a novel multi-domain image analysis dataset, extensive evaluations of several calibration methods across a range of conditions, and recommendations for practitioners to achieve well-calibrated zero-shot image classifiers in diverse contexts.

Introduction

Problem Statement

Recent systems for multimodal representation learning such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023; Alabdulmohsin et al., 2023) are typically pairs of neural networks (an image feature extractor and a text feature extractor) jointly trained on hundreds of millions of `(image, caption)` pairs scraped from the internet. They produce fixed-dimensional numerical representations of their inputs that can be compared, e.g. via cosine similarity. Ideally, the feature vector derived from a given image should have a higher similarity to the feature vector derived from a descriptive caption of that image than to one from an arbitrary text string.

Such systems are used to create binary image classifiers in many domains without requiring the curation of any labeled image data (sometimes called the "zero-shot" setting). For example, a data scientist uses the CLIP text branch to extract the feature vector for the string *"a photo of a dog"*; the values of this vector become the parameters for a `dog` classifier. Then, for each test image, the classifier compares the text feature vector against the feature vector extracted from the CLIP image branch as a subroutine. A similarity above a chosen threshold returns the `dog` label while a similarity below the threshold results in the `not dog` label.

It is straightforward to normalize the cosine scores so that they fall into the range $[0, 1]$ rather than $[-1, 1]$, but there is no reason that this procedure should produce predicted probabilities that match the empirical class distribution. **In this project, we investigate several approaches to produce meaningful probability estimates from CLIP-derived zero-shot image classifiers.** Such approaches include sigmoid scaling (Platt et al., 1999), isotonic regression (Zadrozny and Elkan, 2002), and similarity-binning averaging (Bella et al., 2009). Furthermore, we investigate how well the calibration parameters learned for one zero-shot classifier will work for another zero-shot classifier in the same domain and across domains.

Formally, let $Z : \mathcal{I} \to (-\infty, \infty)$ be a zero-shot classifier, where $\mathcal{I}$ is the space of all input images. For a given image $I \in \mathcal{I}$, $Z(I)$ produces unconstrained scores.

Let $C_\beta : (-\infty, \infty) \to [0, 1]$ be a calibration function parameterized by β that maps raw scores to calibrated probabilities. The composition $C \circ Z$ should produce well-calibrated probabilities, meaning:

$$P(y = 1 \mid C(Z(I)) = p) \approx p \quad \forall p \in [0, 1]$$

where y is the true binary label.

Our goal is to evaluate several such functions C across various subsets of $\mathcal{I}$ to determine if one or a small number of values of β can achieve effective calibration across multiple subsets.

Motivation

There are two primary reasons to explore the calibration of zero-shot classifiers. Firstly, well-calibrated classifiers are generally of interest because they allow end-users to develop more trust in prediction systems, such as by better understanding the level of risk associated with using the predictions for some task (Jiang et al., 2012). Secondly, we are interested in general-use calibrators that work across a variety of classes and domains because they alleviate the need for held-out labeled calibration data. This is particularly salient to the zero-shot classifier setting because there may not be any labeled data at all, let alone labeled data for training a calibrator. We provide guidance as a result of our empirical investigation to allow data scientists to understand not only the effectiveness of various calibration methods for zero-shot image classifiers but also the limits of calibrating them without additional data.

Contributions

The contributions of our work are:

- The curation of a large (100,000+) image dataset representing twenty image categories across two distinct domains, suitable for training and evaluating image classification algorithms (including zero-shot classifiers) and calibrators in challenging settings.
- Extensive empirical results demonstrating the performance of zero-shot classifiers and various calibration algorithms across multiple image classes and domains.
- Evidence-based best practices for data scientists and system builders needing calibrated scores from their zero-shot image classifiers.

Background and Related Work

In machine learning, calibration methods adjust a model's confidence scores to better align with the actual probabilities of an event. This becomes crucial in scenarios where models need to produce probabilities that are both interpretable and actionable. Here, we examine three widely recognized calibration methods – Sigmoid Calibration (also known as Platt Scaling), Isotonic Regression, and Similarity Binning Averaging (SBA) – and discuss their applicability to zero-shot image classifiers.

Sigmoid Calibration is a parametric method that fits a logistic regression model to map a classifier's confidence scores to calibrated probabilities. Proposed by (Platt et al., 1999) for Support Vector Machines, this method minimizes the negative log likelihood using a sigmoid function:

$$P(y = 1|x) = \frac{1}{1 + \exp(A \cdot s(x) + B)}$$

where $s(x)$ represents the raw score from the model, and A and B are learned parameters. This approach is simple to implement and computationally efficient; however, it relies on the assumption that the relationship between scores and probabilities fits a sigmoid curve, which may not be valid for all datasets.

Isotonic Regression is a flexible, non-parametric calibration technique that assumes only a monotonic relationship between predicted scores and actual probabilities. Using the pool-adjacent violators (PAV) algorithm, it constructs a piecewise constant, non-decreasing function to map scores to probabilities. Initially proposed by (Zadrozny and Elkan, 2002), this method is especially useful for handling complex probability distributions. However, it has a tendency to overfit when applied to datasets with limited size or variability.

SBA, introduced by (Bella et al., 2009), is another non-parametric method that groups samples based on their similarity scores into bins based on nearest neighbors. The average observed probability within a bin is used as the calibrated probability for a sample in that group. This method is particularly appealing in high-dimensional applications like zero-shot classifiers, where other calibration methods might struggle to generalize effectively.

There is a growing body of literature on calibration techniques in machine learning. Works such as (Guo et al., 2017) introduced Expected Calibration Error (ECE) as a standard metric for evaluating calibration quality, while (Błasiok and Nakkiran, 2024) introduced smooth ECE (smECE) and recommendations for better calibration reliability diagrams. This work aligns with our zero-shot image classification research as evaluations across diverse conditions are essential to understanding the limits of various calibration methods. Other recent work has also focused on best practices for

calibration, especially of neural network systems, demonstrating the growing importance of this topic (Minderer et al., 2021; Fang et al., 2022; Wagstaff and Dietterich, 2023).

Methods

To build zero-shot classifiers Z, we use the pre-trained `ViT-B/32` architecture from OpenAI's CLIP (Radford et al., 2021) and the pre-trained `ViT-B/16` from Google's SigLIP (Zhai et al., 2023). These are the most computationally efficient choices from each provider, but they still demonstrate strong discriminative performance at the image classification task.

The SigLIP architecture has an interesting feature of its training procedure that is equivalent to sigmoid calibration. The A and B parameters are publicly available as part of the pre-trained parameters, and can be used as a kind of global calibration. We use these sigmoid parameters as part of our evaluation but break them out as a separate category.

For each calibration approach and condition (in-domain, cross-class, and out-of-domain), we evaluate both quantitatively using the smECE metric and qualitatively using reliability diagrams as described in (Błasiok and Nakkiran, 2024).

Specifically, for each choice of image category, calibration method, and feature extractor, we conduct the following procedure:

1. Construct the zero-shot classifier from a CLIP model.
2. Construct a binary (one-vs-rest) version of the in-domain dataset (Places365) by selecting all 5,000 images of the target category and sampling 10,000 images of the non-target category (the ratio of 1:2 is configurable in the code and chosen to balance the class ratio somewhat while still using enough data to get meaningful results).
3. Construct a binary version of the out-of-domain dataset (AID) with a similar sampling process.
4. Divide the in-domain samples into a training partition and a test partition (50/50 split).
5. Train a calibrator on the training partition.
6. Evaluate the calibrator on the in-domain testing partition.
7. Evaluate the calibrator on the out-of-domain testing partition.
8. Save off the calibrator and in-domain testing partitions.
9. Once the calibrators for all classes and in-domain test partitions have been collected, evaluate every calibrator on every other in-domain test partition (to evaluate in-domain cross-class calibration).

Data

For our experiments, we create a custom dataset of 20 different image categories common to the Places365 dataset (Zhou et al., 2017) and AID (Aerial Imagery Dataset) (Xia et al., 2017). This custom dataset contains over 100,000 images allowing sufficient data to split into training and testing partitions for both in-domain and out-of-domain calibration experiments. Examples of Places365 images are shown in Figure 1; examples of AID images are shown in Figure 2.

Figure 1: Example stadium images from Places365

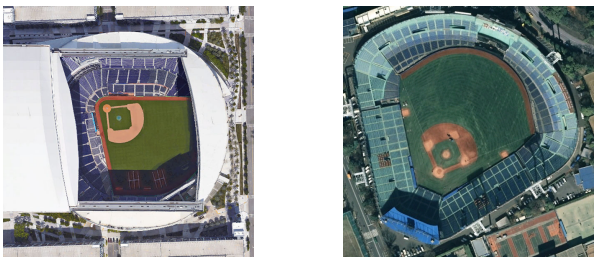

Figure 2: Example stadium images from AID

Results

Our main results are summarized in table 1. In particular, we show that calibration parameters **cannot** be reused across classes or domains without noticeable degradation in calibration quality, regardless of the calibration method used. The implications

Table 1: Comparison of calibration methods across in-domain, out-of-domain, and cross-class settings. Four calibration approaches (Sigmoid, Isotonic Regression, SigLIP, and Similarity-Binning Averaging (SBA)) are evaluated using smoothed Expected Calibration Error (smECE) and accuracy metrics. In-domain testing (including cross-class testing) uses MIT Places 365 data, while out-of-domain testing uses the Aerial Image Dataset (AID). Lower ECE values indicate better calibration, while higher accuracy values indicate better classification performance.

Method	Condition	smECE	Accuracy
Sigmoid	In Domain	0.015	89.9%
	Out of Domain	0.125	85.1%
	Cross Class	0.072	88.3%
Iso. Reg.	In Domain	**0.011**	90.0%
	Out of Domain	0.122	84.4%
	Cross Class	0.071	88.1%
SigLIP	In Domain	0.111	86.0%
	Out of Domain	0.197	76.5%
	Cross Class	0.111	86.0%
SBA	In Domain	0.033	**91.7%**
	Out of Domain	0.127	86.3%

of these findings for practitioners are that in order to achieve meaningful calibration performance, it is essential to curate a calibration dataset and use it to infer a set of task-specific parameters.

A graphical representation of table 1 is given in figure 3.

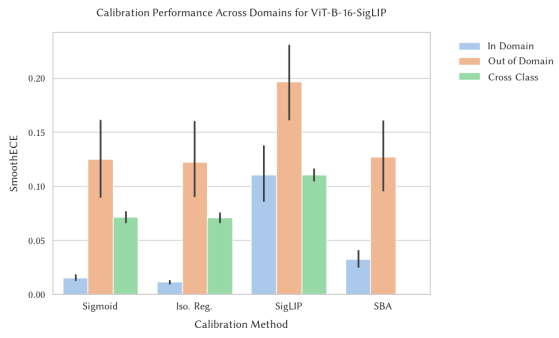

Figure 3: A graphical representation of the jump in calibration errors when moving from in-domain to cross-domain or out-of-domain settings, for all calibration methods (zoom for details)

In order to understand these findings, we investigate the score distributions for several of the tasks. For example, the `river` category in-domain and out-of-domain raw score distributions are shown in figure 4. The target and non-target distributions have significantly different means and skews, meaning that the calibration parameters for simple models like sigmoid calibration have no chance of transferring across domains. This is demonstrated by the associated reliability diagrams, shown in figure 5.

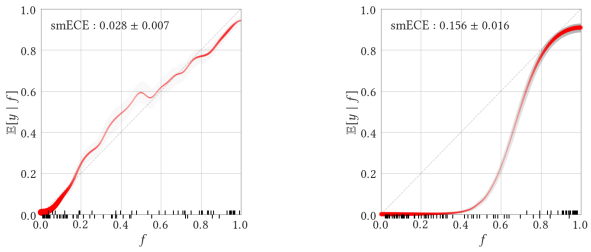

Figure 4: Cosine score distributions from an example class, showing the strong shift between the ground imagery domain and the satellite imagery domain that makes calibration challenging (zoom for details).

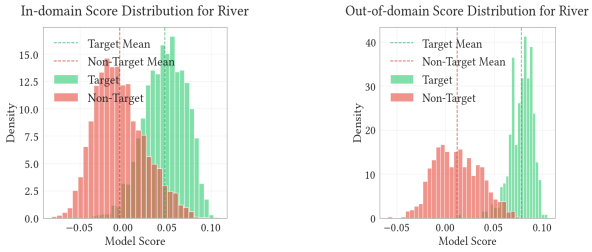

Figure 5: Reliability diagrams representing a strong deviation from ideal calibration in the out-of-domain setting when using parameters trained on the in-domain setting (zoom for details).

As an ancillary finding, our data exploration showed that there is a strong relationship between the accuracy of a classifier and the error rate of a calibrator, shown in figure 6. Specifically, when the accuracy of a classifier is high, it tends to be easier to achieve good calibration performance on it. This suggests that some image categories have high visual saliency that leads to better score discrimination, and in turn, better calibration.

Limitations and Difficulties

A significant challenge encountered in our study was the exceptionally high performance of the CLIP models. These models consistently assigned confidence scores close to 1.0 for correct predictions and near 0.0 for incorrect ones. While this highlights the robust performance of CLIP as a zero-shot classifier, it also made it difficult to effec-

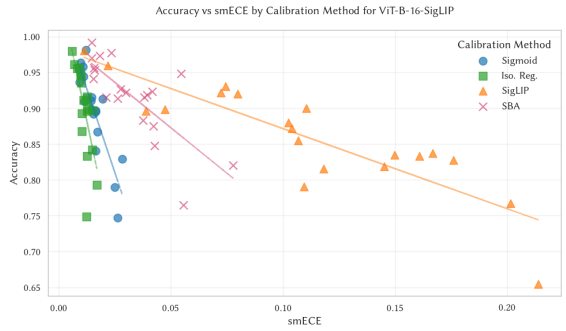

Figure 6: A plot showing the correlation between accuracy and error rate across different calibration methods. Easier classes to discriminate are also easier classes to calibrate (zoom for details).

tively evaluate the performance of the calibration methods. The narrow range of calibration errors observed limited our ability to detect meaningful differences between the methods, necessitating a much larger dataset to identify any significant improvements. This limitation underscores a broader issue: evaluating calibration becomes inherently difficult when a model already demonstrates near-perfect class separability.

Calibration methods rely on the presence of miscalibrated confidence scores to make adjustments, but the high precision of CLIP left minimal room for such corrections. The lack of borderline cases—where confidence scores might be misaligned with true probabilities—further compounded this challenge, particularly when working with smaller or less diverse datasets. While the strong performance of CLIP reduces the need for calibration, it also hinders the assessment of calibration method effectiveness.

To address this issue, we explored utilizing models with lower initial calibration performance as a baseline for evaluating calibration methods. Models that exhibit more overlap in confidence scores between correct and incorrect predictions provide a more challenging context for calibration, allowing for a more detailed comparison of methods. This approach could facilitate a more nuanced assessment of calibration techniques, even in the context of smaller datasets. Ultimately, we addressed the issue by curating a larger, more challenging dataset, but this was very time-consuming.

Future work could explore how the initial calibration of a model influences the effectiveness of calibration methods, examining the interplay between model performance and calibration efficacy. Additionally, this approach could be extended to investigate whether similar challenges arise with other state-of-the-art models across different domains.

Conclusions

Implications

This work enhances the calibration of zero-shot image classifiers, improving their reliability for applications in fields like healthcare, autonomous driving, and disaster response. By analyzing different calibration methods, the study improves confidence estimates, aiding decision-making in high-stakes environments. The transferability of these techniques across classes and domains also highlights their potential for use in diverse, unseen scenarios without requiring extensive retraining. The combination of quantitative metrics (smECE) and qualitative tools (reliability diagrams) sets a benchmark for future calibration research. Furthermore, including out-of-domain experiments addresses the challenge of adapting models to new data distributions, which is crucial in fields like environmental monitoring and remote sensing, where labeled data is scarce.

Future Work

Future work could investigate advanced calibration methods, such as temperature scaling with ensemble models or domain adaptation, and test on more diverse datasets like medical or satellite imagery. Expanding to multiclass classification and studying calibration over time in dynamic environments would deepen insights into complex scenarios and long-term deployment. Exploring calibration across different architectures, improving the interpretability of confidence scores, and applying these methods in real-world settings like species identification or disaster response could enhance practical impact. Additionally, integrating calibration with uncertainty quantification techniques, such as Monte Carlo dropout, could improve handling of uncertainty in predictions.

References

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. In *Advances in Neural Information Processing Systems*, volume 36, pages 16406–16425. Curran Associates, Inc.

Antonio Bella, Cesar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2009. Similarity-binning averaging: a generalisation of binning calibration. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 341–349. Springer.

Jarosław Błasiok and Preetum Nakkiran. 2024. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6216–6234. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. 2012. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Kiri L. Wagstaff and Thomas G Dietterich. 2023. Hidden heterogeneity: When to choose similarity-based calibration. *Transactions on Machine Learning Research*.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.