Strategies to Deal with Unbalanced Data and Uncertainty. Application of Focal Classification

Tina Nguyen, Ricardo Modrego, Vivek Vallurupalli

***Abstract*—This report reviews previous methodologies for improving the performance of an unbalanced data set, with an emphasis on improving calibration and the performance in the least represented classes. Imbalances in data sets pose challenges to machine learning models, necessitating effective strategies for accurate predictions. The importance of addressing this issue lies in its prevalence across various domains, impacting the reliability of classification models. Our strategy focuses on enhancing the calibration of a diatom photo data set through the integration of focal loss and temperature scaling. Our findings show that focal loss in replacement of cross entropy can improve the calibration error of unbalanced data sets. The addition of temperature scaling at varying values also improved the model, but the implementation of focal loss alone yielded the lowest calibration error. Additionally, the upsampling and re-weighting strategies do not produce the anticipated improvements in unbalanced data sets, caused by an overall performance degradation or selective improvement in more unbalanced classes.**

I. Introduction

This project aims to identify and classify microscopic organisms known as diatoms that inhabit aquatic environments. The available data comprises of a long-tailed dataset with a significant disparity in the number of images between certain diatom classes. To tackle this issue, this report focuses on analyzing state-of-the-art strategies for addressing unbalanced data and uncertainty. The goal is to determine the most suitable approaches for diatom recognition and to enhance the reliability of future data analysis.

II. Related Work

A. Focal Classification

The paper On Calibration of Modern Neural Networks [1] and Calibrating Deep Neural Networks using Focal Loss [2] highlight the importance of calibration in modern neural networks. There is a growing prevalence of miscalibration, where the calculated predicted probabilities do not reflect the actual likelihood of the event. In the context of diatoms, certain species have a greater abundance of data than others. Having a calibrated model prevents the underrepresented classification probabilities from being under weighted and underestimated.

On Calibration of Modern Neural Networks [1] applies calibration methods to neural networks for both image classification and document classification tasks. The goal is to evaluate the impact of these calibration methods on the performance of modern neural network architectures and data sets.

For image classification, the study uses six different data sets to train four models. For document classification, the study uses two different datasets, one for Deep Averaging Networks (DANs) and the other for TreeLSTMs. The study trains the models using five different methods: Histogram Binning, Isotonic Regression, Bayesian Binning into Quantiles, Platt Scaling, and Temperature Scaling.

Results from the simpler methods such as temperature scaling and binning were the most successful in calibrating the models. Although temperature scaling introduces only one parameter into the model, it was found to outperform all other methods. While binning did not outperform temperature scaling, the method generally improved calibration. Matrix scaling performed poorly on datasets due to its large number of parameters which tend to overfit with a smaller validation set.

Calibrating Deep Neural Networks using Focal Loss [2] combines temperature calibration with focal loss methods to create calibrated models while preserving accuracy.

Focal loss is a balanced approach to assigning probabilities within models. The loss function assigns weights to loss components based on the model's classification performance, reassigning priorities and adapting to the training set. By encouraging the predicted distribution to have higher entropy (a measure of uncertainty) while minimizing the KL divergence between predicted and true distributions, focal loss can prevent the model from producing overconfident predictions. In contrast, the widely used cross-entropy loss function increases the network's assigned weights when high accuracy is achieved. As a result, the models tend to become overconfident. In focal loss, there is an implicit weight regularization that limits the escalation of the weights.

For image classification, the study uses two different data sets to train on 4 different models. For document classification, the study uses 2 different data sets to train on models, Global Pooling CNN and Tree-LSTM.

The study trains the models using five different methods: Cross-Entropy Loss, MMCE (Maximum Mean Calibration Error), Brier Loss, and LS (Label Smoothing).

Compared to the other methods, focal loss produced the lowest calibration errors with and without temperature scaling. Focal loss also showed inherent calibration as the combination of temperature scaling and focal loss yielded lower temperatures around 0.9 to 1.1 while the other methods yielded temperatures between 2.0 and 3.0. After applying temperature scaling to focal loss, the model also became more confident in its correct predictions.

B. Unbalanced Data

The paper Deep Long-Tailed Learning: A Survey [3] analyzes numerous studies that have been conducted to address the problem of deep long-tailed learning. Existing studies in deep long-tailed learning are classified into three main categories: Class Re-balancing, Information Augmentation, and Module Improvement. These categories are distinguished by the way of addressing the unbalanced problem.

In the first case, **class re-balancing** methods aim to re-balance the class distribution during training to ensure that tail classes receive more emphasis. These methods can be implemented using the following approaches:

- Re-sampling focuses on adjusting the number of samples of each class for the model training.
- Class-sensitive Learning seeks to particularly adjust the training loss values for various classes to re-balance the uneven training effects caused by the imbalance issue.
- Logit Adjustment seeks to resolve the class imbalance by adjusting the prediction logits of a class-biased deep model.

Regarding **information augmentation**, it focuses on strategies that use various data augmentation techniques to artificially increase the number of instances available for tail classes. It has two different approaches:

- Transfer learning seeks to transfer the knowledge from a source domain (e.g., datasets) to enhance model training on a target domain.
- Data Augmentation aims to enhance the size and quality of datasets by applying pre-defined transformations to each data/feature for model training.

Concerning **module improvement**, it focuses on improving different components of the neural network architecture. These methods are divided into four categories.

- Representation learning: improves representation learning at the feature level.
- Classifier design: enhances the model classifier normalizing sample's weights and features.
- Decoupled training: aims to boost the learning of both the feature extractor and the classifier, separating its training.
- Ensemble learning: improves the whole architecture generating and combining multiple network modules.

Empiric studies have provided key findings and relevant conclusions. Firstly, the results demonstrated that most long-tailed learning methods outperformed the Softmax baseline in terms of accuracy, highlighting the effectiveness of these specialized techniques. Upper Reference Accuracy and Relative Accuracy were introduced as evaluation metrics. While UA remained the same as the baseline for most methods, RA provided a more comprehensive assessment by considering factors beyond class imbalance.

Additionally, it was observed that combining Class Re-balancing techniques without careful design did not necessarily lead to better performance. Also, Transfer-Based Methods were found to be beneficial when combined with other long-tailed learning paradigms. They complemented class re-balancing and module improvement methods. Lastly, data augmentation methods like RandAugment, consistently improved the performance of various long-tailed learning paradigms.

C. *Conclusion*

After a thorough analysis of the current literature, we have identified the most fitting methods for addressing unbalanced data and uncertainty within the constraints of time and our existing knowledge for this project. These methods include Re-Sampling, Re-Weighting, Data Augmentation, and Representation Learning. To tackle the problem during the training and inference phase, we also plan to apply the methods of Focal Loss paired with Temperature Scaling. These approaches will be instrumental in handling the data imbalance effectively and enhancing the diatom recognition process.

III. Methodology

A. *Focal Classification*

This paper proposes using focal loss as the main loss function and temperature scaling for post-processing to address the uncertainty issue of our data set. Currently, the algorithm is using cross-entropy as the loss function. Whilst this is a popular method to address the uncertainty issue, focal loss can offer a more effective solution by down-weighting abundant examples. Additionally, incorporating temperature scaling allows us to calibrate the model's confidence scores, further ensuring accurate and reliable predictions.

To implement the focal loss function, PyTorch's simple loss function [4] will be integrated into the model. For this solution, the focal loss with only the gamma parameter will be utilized. The gamma parameter will be 2.0, matching the one used in Calibrating Deep Neural Networks using Focal Loss [2]. This will balance the classification of the data set's abundant and rare diatoms without the need to include weight parameters.

To determine if the proposed solution will provide better results, calibration plots will be generated from the predicted probabilities. The x-axis will represent the average predicted confidence within each bin while the y-axis will represent the observed accuracy within each bin. This is the fraction of examples in each bin that were correctly classified. If the model is well-calibrated, the predicted probabilities should be similar to the actual accuracy and the observed accuracy should be similar to the predicted confidence.

The expected calibration error will also be calculated by comparing the predicted confidence scores with the observed accuracy. The algorithm will compute the absolute difference between the predicted confidence and accuracy and will be averaged. To obtain the ECE, the result is further divided by the total number of predictions.

Both the plots and the calibration error values will be compared to determine which focal loss function will provide the best results. A lower calibration error value shows that the predicted probabilities are closer to the accuracy. For the plots, the calibration curves will be compared to determine which result aligns with the ideal diagonal line, representing perfect calibration.

To ensure that the accuracy of the model is similar while using cross-entropy and focal loss, the f-scores will be computed and compared.

B. Unbalanced Data

To enhance performance when dealing with unbalanced datasets, various methods outlined in the literature survey will be implemented on the original diatoms dataset, and their effectiveness will be assessed.

Firstly, the Oxford-IIIT Pet Dataset [5] will be utilized to evaluate the behavior of the current neural network on a balanced dataset. This dataset is a collection of images featuring various categories of pets, primarily focusing on different breeds of cats and dogs, including images of 37 different pet categories. The dataset will be intentionally made progressively more unbalanced by removing samples from some classes to observe changes in performance. The data balancing techniques described in section **??**, Re-Sampling and Re-Weighting, will then be applied to both the Oxford-IIIT Pet Dataset and the larger, more complex diatoms dataset.

The selected evaluation metrics for assessing the performance of the techniques are the F-score and Accuracy, with a particular emphasis on the F-Score. This choice is motivated by the F-Score's ability to consider both false positives and false negatives in predictions, providing a more realistic performance assessment when dealing with unbalanced data sets.

Re-sampling techniques primarily comprise up-sampling and down-sampling. Up-sampling aims to balance the number of samples for each class by duplicating images for classes with fewer samples until they match the class with the most samples. Down-sampling follows a similar principle but reduces the number of samples in classes with more data to match those with the fewest samples. Up-sampling will be implemented as it tends to yield better results without information loss [3].

For re-weighting, different weights will be assigned to the classes during neural network training. The weights will be assigned to be inversely proportional to the frequency of each class, giving higher weight to images of classes with fewer samples, resulting in an equal weight for each class as a whole.

The data augmentation module is already integrated into the code [6].

It is important to note that the primary challenge in implementing this methodology is that it is not a direct replication of an existing paper; rather, it represents an approach developed by the authors. Additionally, applying these techniques to the diatoms dataset is resource-intensive due to the large number of images, especially in the class with the most samples. This results in extended training times, increased complexity, and potential data volume issues.

IV. Results

A. Focal Classification

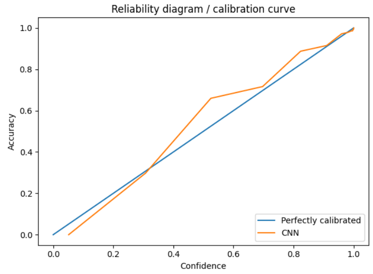

Fig. 1. The calibration plot comparing the performance of the model trained with focal loss to the perfectly calibrated line.

Fig. 1 shows the calibration plot when cross-entropy was applied, showing a calibration error of 0.02016. When cross-entropy was replaced with focal loss, we received a calibration error of 0.015173 which improved the model's calibration.

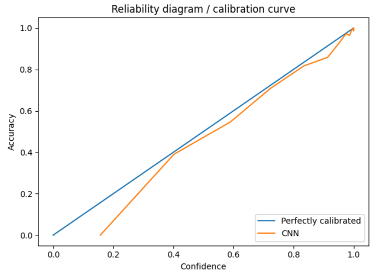

Fig. 2. The calibration plot comparing the performance of the model trained with focal loss to the perfectly calibrated line.

In Fig. 2, the calibration plot demonstrated a much higher calibration error of 0.21639 as compared to the performance achieved with cross-entropy.

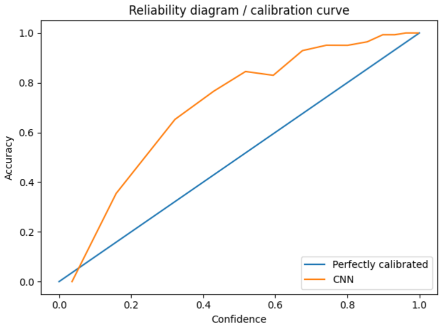

Fig. 3. The calibration plot comparing the performance of the model trained with focal loss to the perfectly calibrated line. Additionally, temperature scaling was applied post-training.

In contrast, Fig. 3 shows the calibration error focal loss combined with temperature scaling. From previous research, it was expected that temperature scaling, applied post-training, would align the predicted probability with the actual probability.

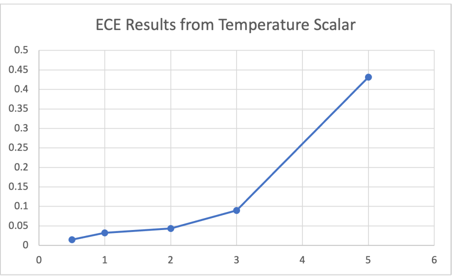

Fig. 4. Temperature scalars 0.5, 1, 2, 3, and 5 and the relative calibration errors that were tested on the model.

Higher temperatures that are greater than 1 were used to smooth out the probabilities, making them more uniform. Fig. 4 shows higher temperature scalars that were tested on the model to determine which scalars would best optimize the performance. Surprisingly, the lower scalars slightly improved the calibration error. 0.05 was the scalar that yielded the lowest error but still failed to outperform focal loss alone.

B. Unbalance Data

1) Upsampling: In this section, the Oxford-IIIT Pet Dataset [5] is intentionally made unbalanced. An unbalancing function is applied, reducing the number of samples within each class exponentially for various unbalance parameter values, ranging from 0.0 to 0.12. A value of 0.0 implies no unbalance, while 0.12 results in the class with the fewest samples having only one.

$$MF = (1 - Up)^i \tag{1}$$

The formula for the Multiplication Factor (MF) is represented in the Eq. 1, where *Up* corresponds to the value of the unbalance parameter and *i* stands for the specific class that is being unbalanced.

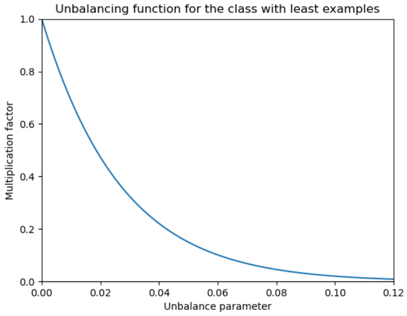

Fig. 5. Unbalancing function applied to the class made the most unbalanced.

Figure 5 illustrates the multiplication factor applied to the number of samples in the most unbalanced class, with a designated value of 37 for variable *i*.

The F-Score metric analysis is conducted on the Oxford-IIIT Pet Dataset without any technique for handling unbalanced data, as well as with the application of an upsampling strategy. The goal is to evaluate the performance of the upsampling strategy.

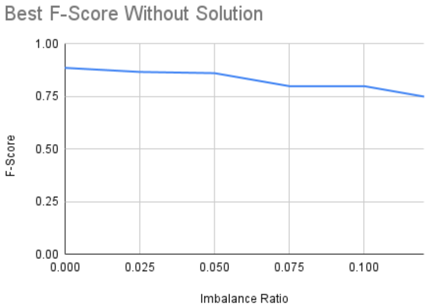

Fig. 6. Best F-Score obtained in 100 epochs of training for the different values of the Unbalance parameter, without any strategy implemented to deal with unbalanced data.

Fig. 6 illustrates the best F-Score obtained over 100 training epochs for different unbalance parameter values, 0.0, 0.025, 0.05, 0.075, 0.1 and 0.12. This information is important to analyze the performance of the upsampling solution and prove whether it provides better results. The obtained results show the anticipated trend where predictions deteriorate with a more unbalanced dataset.

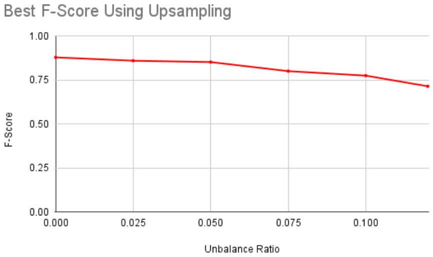

Fig. 7. Best F-Score obtained in 100 epochs of training for the different values of the Unbalance parameter, with the upsampling strategy implemented.

Fig. 7 showcases the best F-Score obtained under the exact conditions as Fig. 6, but with the implementation of the upsampling strategy. It reveals a similar trend, with the F-Score inversely proportional to the unbalance parameter.

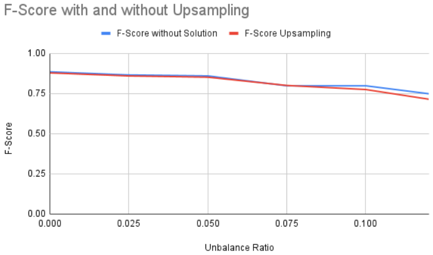

Fig. 8. Comparison between the best F-Score obtained in 100 epochs of training for the different values of the Unbalance parameter, with and without the upsampling strategy implemented.

Fig. 8 compares the best F-Score obtained in 100 epochs with and without the upsampling strategy. It has been included to analyze whether a better result was obtained using the upsampling technique. Surprisingly, the upsampling solution doesn't yield any improvement in the Oxford-IIIT Pet Dataset. This may be due to either worse performance across all classes or an improvement in more unbalanced classes but worse performance in classes with higher sample numbers. This could be an interesting result, as the aim is to improve the performance in the least represented classes.

After analyzing the Oxford-IIIT Pet Dataset, the upsampling technique is applied to the Atlas diatoms dataset that motivated this project.

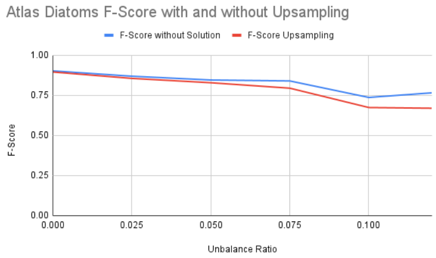

Fig. 9. Comparison between the best F-Score obtained in 100 epochs of training for the different values of the Unbalance parameter, with and without the upsampling strategy implemented, applied in the Atlas diatoms dataset.

Fig. 9 compares the best F-Score obtained in 100 epochs with and without the upsampling strategy applied to the Atlas diatoms dataset. Similar to the Oxford-IIIT Pet Dataset, the F-Score for different unbalance parameter values is slightly lower with the upsampling technique. This could be due to worse performance across all classes or improvement only in more unbalanced classes, leading to an overall performance decline. As previously stated, this result is intriguing, because the objective is to enhance performance in the least represented classes.

2) Re-weighting: Similarly to the upsampling section, the same Oxford-IIIT Pet Dataset [5] is used for the re-weighting section, as well as the same unbalancing function and values in order to allow for accurate comparisons in both strategies.

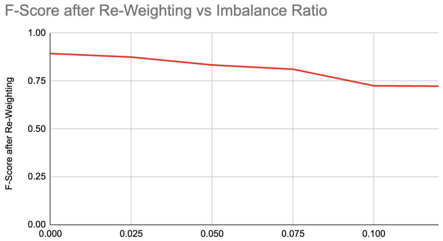

Fig. 10. Best F-Score obtained in 100 epochs of training for the different values of the Unbalance parameter with the Re-Weighting strategy.

Fig. 10 displays the highest F-Score achieved after 100 epochs of training with the re-weighting strategy against each of the different unbalance parameters (0, 0.025, 0.05, 0.075, 0.1, 0.12).

Fig. 11 compares the graph of the best F-Scores for each unbalance parameter with and without re-weighting. It is clear that both graphs are extremely similar. Similarly to the upsampling graph, there is no clear improvement with the re-weighting, with there only being any improvement at all with unbalance parameters of 0, 0.025, and 0.075.

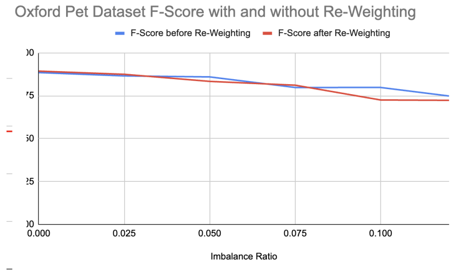

Fig. 11. Both the best F-Score obtained in 100 epochs of training for the different values of the Unbalance parameter with the Re-Weighting strategy and with no strategy.

The results of this graph can lead to questions about why the re-weighting strategy – one supposed to help with model performance with increasing unbalance – wasn't able to make a significant difference at all. Regardless, it is a result worth taking note of.

V. Conclusion

In conclusion, this project undertakes an exploration of methods to enhance the performance of a model when dealing with unbalanced datasets, with a specific focus on a diatoms dataset. The Oxford-IIIT Pet Dataset serves as preliminary testing grounds, providing insights into the behavior of the neural network on a balanced data set and its response to intentional unbalancing. Furthermore, the exploration of focal loss, coupled with temperature scaling, on the diatoms data set showcases enhancements in the model's performance.

The chosen evaluation metrics, F-Score, Accuracy, and Calibration Error guide the assessment of various techniques outlined in the literature survey. Among these, Re-Sampling, Re-Weighting, and Data Augmentation emerge as the most promising solutions.

One part of the study involves the intentional unbalancing of the Oxford-IIIT Pet Dataset using an exponential function. The subsequent analysis, particularly employing the upsampling and re-weighting strategies, provides valuable insights into the challenges and potential improvements when faced with unbalanced data. Another part of the study focuses on enhancing the diatoms data set by implementing focal loss and temperature scaling, and further fine-tuning the gamma parameter to achieve the best calibration error.

The Oxford-III Pet Dataset results, illustrated through figures, demonstrate that while the upsampling technique follows an expected trend, it surprisingly does not yield improvement in the Oxford-IIIT Pet Dataset. Further, the application of the upsampling technique to the larger and more complex Atlas diatoms dataset reflects a similar trend, with a slightly lower F-Score for different unbalanced parameter values. This outcome raises intriguing questions about the nuanced impact of the upsampling strategy, especially concerning its effectiveness in enhancing the performance of the least represented classes. As for the re-weighting ,while for some unbalance parameters there was some marginal improvement in F-Score, there was not an overall improvement in performance. Similarly to the upsampling technique, the performance of the re-weighting strategy raises questions about it's viability.

In turn, the diatoms data set showed promising results for the usage of focal loss over cross entropy for unbalanced data sets. While the addition of temperature scaling did not achieve the expected results, further testing and tuning of the temperature parameter may be needed to determine the most optimal scalar. More research on the loss function's characteristics and potential overfitting concerns are required to understand why the combination did not achieve the best results.

Regarding future directions, a comprehensive examination of performance on a per-class basis is essential to assess the impact of the upsampling technique. This analysis will help determine whether the method effectively enhances the performance of classes with fewer samples or if its utility is limited. Additionally, it is advisable to explore other pertinent techniques, such as representation learning, which were examined in section II-C. Implementing these approaches holds promise as they may contribute significantly to improving overall model performance.

REFERENCES

[1] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating deep neural networks using focal loss," 09 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf

[2] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger, "On calibration of modern neural networks," 08 2017, accessed on 2023-09-20. [Online]. Available: https://arxiv.org/pdf/1706.04599.pdf

[3] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," 10 2021. [Online]. Available: http://arxiv.org/abs/2110.04596

[4] "Focal loss torch," https://pypi.org/project/focal-loss-torch/1, accessed on 2023-10-11. Released on 2023-04-24.

[5] monkeydoodle@gmail.com, "The oxford-iiit pet dataset dataset," https://universe.roboflow.com/monkeydoodle-gmail-com/the-oxford-iiit-pet-dataset, may 2022, visited on 2023-11-27. [Online]. Available: https://universe.roboflow.com/monkeydoodle-gmail-com/the-oxford-iiit-pet-dataset

[6] A. Venkataramanan, "Data loader," https://gitlab.georgiatech-metz.fr/sp/uncertain-unbalanced-data/tree/master/dataset_utils, 2023.