Improving Strong Gravitational Lensing Detection with Active Learning and Tailored Loss Functions Under Data Scarcity*

April 15, 2025

Strong gravitational lensing is a rare yet critical phenomenon in astrophysics, offering insights into dark matter, galaxy evolution, and cosmology. However, its scarcity in observational data presents a significant challenge for machine learning models, particularly in supervised learning frameworks that rely on large amounts of labeled examples. In this study, we propose an approach that integrates active learning with specialized loss functions to improve the detection of strong lensing events under limited data conditions. We employ the BADGE active learning strategy to select the most informative samples for labeling and evaluate two custom loss functions—Focal Loss and RankNet Loss—against a baseline model using standard cross-entropy. Our experiments reveal that while Focal Loss shows modest improvements in surfacing true lensing candidates, its computational cost may limit practical application, especially in active learning contexts. RankNet Loss, on the other hand, suffers from pairing limitations due to class imbalance. We also introduce a more task-aligned evaluation metric based on the number of correctly identified lensing cases in the top 10% of ranked predictions. While no single method decisively outperforms the others, our results offer useful guidance for balancing performance and efficiency in rare-event classification problems and lay the groundwork for future enhancements through better feature learning, metric design, and data acquisition strategies.

Table of contents

*Code and data are available at: https://github.com/Junbo345/Actinve_learning_Strong_Lensing.

1 Introduction

Strong gravitational lensing occurs when a massive foreground object, such as a galaxy cluster, bends the light from a more distant source, producing distinctive arcs or rings in astronomical images. Although these phenomena are of great importance to cosmology, they are extremely rare—accounting for only a small fraction of all astronomical observations. Traditional supervised learning approaches typically require large amounts of labeled data, which is not feasible in this context due to the scarcity of strong lensing examples.

To overcome this challenge, we adopt a state-of-the-art active learning strategy known as BADGE, which selects the most informative data points for human annotation. This method aims to improve training efficiency by prioritizing samples that are both uncertain and diverse (Ash et al. 2020). BADGE has shown strong performance in other low-label settings and is recognized as one of the most effective techniques in active learning (Ash et al. 2020).

Despite this, we still face a significant issue: class imbalance. In our dataset, strong lensing instances represent only about 8% of the total data. This imbalance poses difficulties for both model training and evaluation. To address this, we integrate loss functions and evaluation metrics specifically designed for imbalanced classification tasks.

The remainder of this paper is structured as follows: Section Section 2 outlines our approach, including the active learning framework, loss functions, and evaluation metrics used. Section **?@sec-results** presents the performance results under the proposed method. Finally, Section **?@sec-diss** discusses the implications of our findings, current limitations, and directions for future research in detecting strong lensing events.

2 Method

To evaluate the effectiveness of the proposed loss functions and evaluation metric, we use available strong lensing data in a regression framework. Features are first extracted using a convolutional neural network (CNN), and then concatenated with expert-provided labels for training. The resulting dataset contains approximately ~ total samples, with ~ representing strong lensing cases.

2.1 New loss function

We explore two specialized loss functions: RankNet loss and Focal loss.

2.1.1 RankNet Loss:

RankNet loss is designed to capture the relative ordering between data points based on their labels. Given training data points $x_1, x_2, \ldots, x_n$ with corresponding rankings $y_1, y_2, \ldots, y_n$, we consider all ordered pairs (x_i, x_j) where $y_i \neq y_j$. The loss is computed as:

$$L = \sum_{i<j} -y_{ij} \log P(i > j) - (1 - y_{ij}) \log(1 - P(i > j))$$

Here, $y_{ij} = 1$ if x_i is ranked higher than x_j, and 0 otherwise (assuming no ties in ranking). The probability that x_i should be ranked above x_j is defined as:

$$P(i > j) = \frac{1}{1 + e^{-(s(x_i) - s(x_j))}}$$

. where $s(\cdot)$ is the model's predicted ranking score. This loss penalizes the model when it incorrectly ranks a more relevant (higher-lensed) sample lower than a less relevant one.

2.1.2 Focal Loss:

Focal loss is particularly effective in handling class imbalance by focusing learning on hard-to-classify examples. It is defined as:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

Here, p_t is the predicted probability for the true class: the model's predicted probability if the true label is strong lensing, and $1 - p_t$ otherwise. The weighting factor α_t is set to 0.75 for strong lensing samples and 0.25 for non-lensing ones. This loss down-weights well-classified examples and puts more emphasis on correcting misclassified strong lensing instances, which are more valuable in our context.

2.2 New evaluating matrics

Due to the severe class imbalance—strong lensing events account for only ~8% of the dataset—traditional evaluation metrics like overall accuracy can be misleading. For example, a model that predicts all samples as non-lensing would achieve high accuracy, yet fail to serve our scientific objectives.

To better reflect the model's practical performance, we propose a more intuitive and goal-aligned metric: the number of correctly identified strong lensing samples within the top 10% of the highest-ranked predictions.

This metric directly measures the model's ability to surface valuable candidates for human review and is more consistent with the real-world usage of the system—prioritizing rare but scientifically significant events for expert validation.

3 Results

To evaluate the two proposed loss functions using the new metric, we built a simple neural network in PyTorch. The model architecture consists of an input layer with 40 neurons, one hidden layer with 128 neurons, and a single output neuron representing the predicted score of the input sample. The dataset was split into training and testing sets with a 7:3 ratio. To improve training efficiency, we used mini-batches of size 128 and set the number of training epochs to 100 to avoid overfitting based on preliminary fine-tuning.

For comparison, we also trained a baseline model using standard cross-entropy loss. All models were trained and evaluated across 10 independent runs using different random seeds.

Figure 1 presents the average number of correctly identified strong lensing instances among the top 100 ranked predictions (approximately the top 10% of the test set). This metric directly reflects the model's ability to surface the most relevant candidates.

We observe that all three loss functions—cross-entropy (MLP), RankNet, and Focal Loss—perform similarly, with the average number of correct predictions hovering around 40. While RankNet does not show clear improvement over the baseline, Focal Loss slightly outperforms cross-entropy in some runs. However, the differences are not statistically significant, as indicated by the overlapping error bars.

Overall, while Focal Loss shows some promise, none of the alternative loss functions demonstrated a substantial performance gain under this evaluation setting.

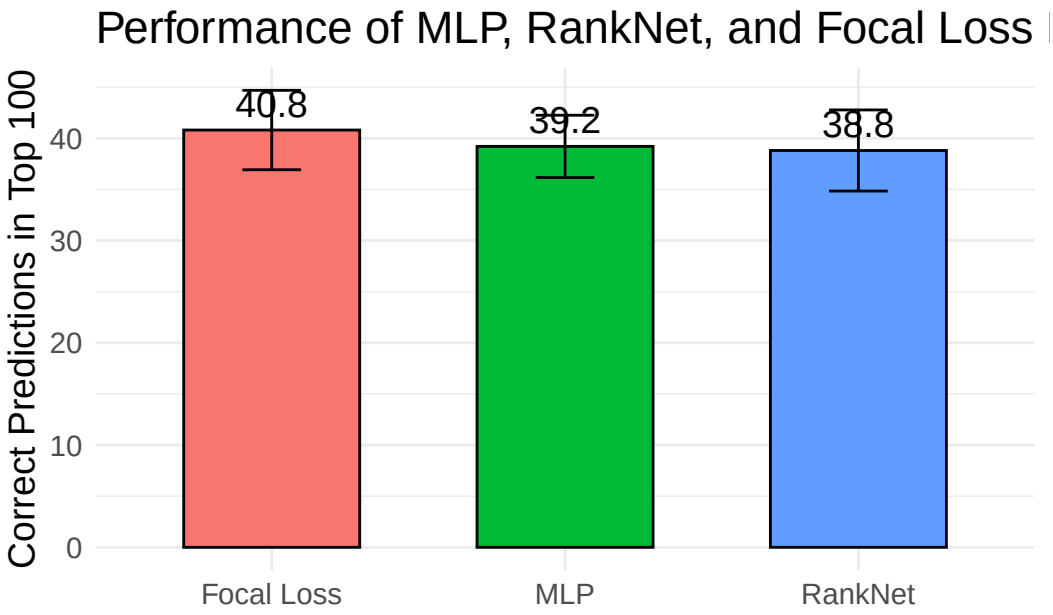

Figure 1: Performance of two new proposed loss functions compared to crossentropy loss. They are evaluated based on number of correct labels in the top 100 labels predicted by the neuronetwork, roughly 10% of all the testing data. We see that rank loss does not outperform crossentropy loss while focal loss did but there is no promising beatens since the error bars contain each other's mean.

4 Dsicussion

4.1 Conslusion and limitation

Based on our experiments, Focal Loss appears to be the most promising option among the three loss functions tested. However, in practice, it introduces increased computational complexity during training—particularly when used in combination with active learning strategies like BADGE, which requires computing gradients with respect to the penultimate layer of the neural network.

With Cross-Entropy Loss, the gradient has a simple, closed-form expression, which makes it computationally efficient and easier to integrate into BADGE. In contrast, Focal Loss includes additional terms involving α_t and γ, which complicate gradient calculations and increase training time. To make Focal Loss more viable in practice, it may be necessary to fine-tune its hyperparameters to improve both performance and efficiency.

As for RankNet Loss, its application in our context is limited by the rarity of training pairs with differing labels in each mini-batch. Since strong lensing cases are so scarce, many batches end up containing only one class, making it impossible to form valid comparison pairs. This severely hampers the effectiveness of the RankNet approach under class-imbalanced conditions.

4.2 Future work

While our current study highlights some key findings, it also reveals several opportunities for further research and development. One promising direction is to fine-tune the CNN feature extractor used to generate input representations for the model. In our experiments, the CNN was kept fixed, but optimizing it alongside the main model could lead to more expressive features and improved performance in identifying strong lensing events.

Another avenue is the exploration of alternative loss functions or ranking strategies that are more suitable for highly imbalanced datasets. For example, newer ranking-based methods or hybrid loss formulations might offer better trade-offs between learning signal and computational complexity, especially when valid data pairs are scarce—as was the case with RankNet in our experiments.

In addition, refining the evaluation metric to better reflect real-world objectives would strengthen the connection between model outputs and practical use. While our current metric—correct predictions in the top 10%—is more aligned with human review processes, future work could explore even more nuanced measures of model utility in scientific workflows.

Finally, incorporating advanced active learning strategies represents a critical path forward. By intelligently selecting which samples to label, we can increase the diversity and informativeness

of the training data while minimizing labeling effort. This is especially relevant given the scarcity of strong lensing cases and the high cost of expert annotation.

Together, these directions could significantly enhance the robustness, accuracy, and real-world applicability of models developed for detecting strong gravitational lensing events.

References

Ash, Jordan T., Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. "Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds." In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1906.03671.