Team Discount GPT: Large Binocular Telescope Misalignment Project Midterm Report

Ryan Luu[1], Bowman Brown[1], Alfianto Widodo[1], Mahmoudreza Dehghan[1], and Deqing Fu[1]

[1]Department of Computer Science, University of Southern California, 941 Bloom Walk, Los Angeles, CA 90089 , USA

1. CURRENT DATASET

The objective of our research is to develop a preliminary model capable of identifying misalignment information in an optical system by analyzing the Point Spread Function (PSF) subsystem. However, due to the intricate nature of the data parameters and the current scope of the project, we will utilize synthetically generated data with a reduced scope, as opposed to real-world data. For the purpose of our investigations, we will consider a simplified PSF, defined as a two-dimensional (2D) image that represents the spot shape generated by the optical system. The input parameters for our algorithm comprise positional and tilt coordinates (d_x, d_y, d_z, t_x, t_y), PSF position (p_x, p_y), field coordinates (field_x, field_y), and the image. Given p_x, p_y, and the image as input, our model will try to predict the misalignment information d_x, d_y, d_z, t_x, t_y.

Table 1. Example Line from CSV of Generated PSF Dataset

	d_x	**d_y**	**d_z**	**t_x**	**t_y**	**p_x**	**p_y**	**field_x**	**field_y**	**data_img**
0	-2.0	-2.0	-2.0	0.0	0.0	52.026...	51.736...	0.104	-0.104	”[[0. 0. ... 0. 0.]

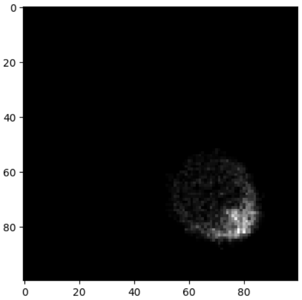

Figure 1. Image of Ray-Traced Generated Star Image.

2. WHAT HAS BEEN DONE SO FAR

2.1 Loss functions

In our preliminary experiments, we have employed the Mean Squared Error (MSE) as the loss function for our baseline models. However, we acknowledge that MSE may be less resilient to the presence of outliers in our training dataset. To address this limitation, our research aims to explore alternative loss functions that can facilitate more accurate regression analysis, while maintaining robustness against outliers. One potential candidate is the Huber loss, which amalgamates the advantageous properties of both Mean Absolute Error (MAE) and MSE. The mathematical formulation of the Huber loss can be found in the following equation:

$$L_\delta = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & if\ |(y-\hat{y})| < \delta \\ \delta((y-\hat{y}) - \frac{1}{2}\delta) & otherwise \end{cases}$$

Huber loss applies a quadratic penalty when the difference between the true and predicted value is less than the hyperparameter delta. Otherwise, it applies the absolute value of the difference as penalty. Our research aims to explore and potentially combine variations of the standard Huber loss as presented in the following studies: ”Generalized Huber Loss”[1] and ”Alternative Probabilistic Interpretation of the Huber Loss”.[2]

The paper on Generalized Huber Loss proposes an algorithm to efficiently find the parameters that lead to minimizing the loss for robust learning and efficient minimization. The paper on Alternative Probabilistic Interpretation of the Huber Loss relates minimizing the loss to minimizing an upper-bound on the Kullback-Leibler divergence between Laplace distributions, where one distribution represents the noise in the ground-truth and the other represents the noise in the prediction. This approach is also tested on vision tasks such as object detection with RetinaNet[3] and R-CNN[4] which goes along with the objective of this research.

2.2 Deep Learning Models

2.2.1 Regression algorithms

We employed fundamental regression algorithms without the use of the image as baseline models, given the nature of the problem. The Large Binocular Telescope Misalignment issue presents a unique and intricate challenge, and we anticipated that the performance of these basic models would be constrained. Nevertheless, our objective was to investigate the behavior of these models on our dataset to gain insights into their applicability and limitations.

As mentioned in section 2.1, we used Mean Squared Error (MSE) as our evaluation metric for the regression algorithms. The best MSE score was obtained by the Random Forest algorithm with a value of 0.484, followed by Gradient Boosting with a score of 0.483, Linear Regression with a score of 0.429, and KNN with a score of 0.400, as shown in figure 2. MSE scores confirm our expectation that these simple models could not explain a significant amount of the variance in the target variable.

One potential explanation for the restricted performance of these models is their limited capacity to capture the various factors contributing to the misalignment. Furthermore, the dataset may have exhibited noise and high-dimensionality, posing challenges for these models to effectively address with just regression alone. To mitigate these limitations, it is essential to investigate the more sophisticated algorithms outlined in the subsequent sections. In summary, although the results were anticipated, examining the performance of these foundational models offers a valuable baseline for future comparisons and a starting point for enhancing prediction accuracy.

2.2.2 Convolutional Neural Network

In order to establish a baseline for our research, we opted to employ a Convolutional Neural Network (CNN) as our initial approach, given its proven efficacy in image processing tasks. Despite the limited success of CNNs in high-precision regression domains, we believe that utilizing this architecture as a foundation for our project could provide valuable insights and pave the way for future refinements.

The CNN had a very simple structure of $conv1 \rightarrow pool \rightarrow conv2 \rightarrow fc1 \rightarrow fc2$ where the last layer would output our 5 misalignment parameters. Training was performed with Early Stopping in order to account for overfitting and we managed to acheive a MSE Score of 0.9741. Results of training can found in figure 2 below.

2.2.3 Vision Transformer

We used the Hugging Face Transformers library to implement the Vision Transformer (ViT). Since the ViT has a large computational cost, we first ran truncated experiments to test and compare potential configurations before selecting a configuration to run. The model was trained for 100 epochs with a learning rate of 0.001 configurations with the Adam optimizer.

We experimented with both using the pre-trained weights provided by Hugging Face as well as and random weight initialization to compare results. Based on a truncated experiment of 20 trials, we selected the pre-trained Hugging Face weights. We note that more in-depth experimentation is required to determine if the pre-trained weights improve the final results.

Our data is unique in that it is multi-modal, and as such, there are many possible ways to combine the input information. Specifically, we have a 100x100 greyscale image and two input scalars px and py. Our current experiments with ViT involve a simple combination of the image and the scalar values. First, we use a linear

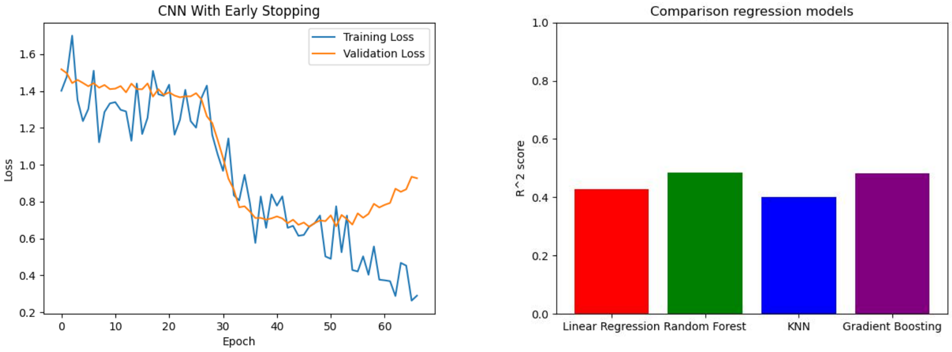

Figure 2. Results of first CNN Trial (left) and Basic Regression Models (right). For the CNN, the Mean Squared Error (MSE) = 0.9741

layer to map each px and py value into a 100x100 map. The two maps are then concatenated with the input image along the channel dimension. This creates a 3-channel image that can be fed into the vision transformer. After 100 epochs, we obtained an MSE score of 1.446 on the validation dataset.

2.2.4 Resnet-18 Ensemble

As part of our initial experimentation, we attempted to form an ensemble of simple models for our regression task. We used a total of 20 Resnet-18 models in the ensemble. Each Resnet-18 model was trained with a learning rate of 0.01 and the Adam optimizer. To combine the results of our ensemble models, we used mean averaging. Each model predicted 5 real valued output scalars, allowing us to use averaging as a simple approach that considered each model as having equal contribution to the final results. After 100 epochs, we obtained an MSE score of 2.537 on the validation dataset.

2.2.5 Resnet-MLP

Our third approach attempted to facilitate the combination of our multi-modal image and scalar value inputs. We used a Resnet-50 to encode each image into a vector of length 512, which was then concatenated with the two px and py scalar values to become a vector of length 514. The 514-length vector was then fed into a six layer MLP that used ReLU activation functions between layers.

Our approach processed the image data separately from the scalar inputs to facilitate the combination of the multi-modal data. The proposed architecture took advantage of the CNN's ability to work well with image data, which has been reported in previous work. After 100 epochs, we obtained an MSE score of 4.110 on the validation dataset.

3. PLANS FOR NEXT ITERATION OF WORK

Going forward, new approaches to combining the image and scalar input information need to be explored. Our initial analysis had focused on deep learning approaches for regression. However, given the multi-modal nature of our input data, more advanced techniques for data combination may provide significant improvements to our results. The more advanced Deep Learning models have not yet shown good results, likely due to issues with tuning or implementation.

Additionally, more fine-tuning of hyperparameters may be required. Our work so far has consisted of prototyping models and implementing baselines from previous work. Results may be strongly influenced by the hyperparameter selection, and as such, tuning hyperparameters is an important step in delivering useful models.

With regard to the dataset size, we currently have only 675 examples at our disposal, primarily due to the time-consuming nature of generating and raytracing them. Nonetheless, efforts are underway to increase this number. This expansion can be achieved by increasing the variations between each dependent variable in the MATLAB program, though meticulous attention to detail is required to ensure that the bounds of each variable remain within acceptable limits.

REFERENCES

[1] Gokcesu, K. and Gokcesu, H., "Generalized huber loss for robust learning and its efficient minimization for a robust statistics," (2021).

[2] Meyer, G. P., "An alternative probabilistic interpretation of the huber loss," (2020).

[3] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," (2018).

[4] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," (2014).

4. CONTRIBUTIONS

Everyone so far has been contributing to both the process of this research project as well as to the current Midterm report here.

Ryan Luu - Worked on Generating the Dataset and the CNN Model.

Bowman Brown - Worked on Vision Transformer, Resnet-18 Ensemble, and the Resnet-MLP

Alfianto Widodo - Worked on researching the loss functions.

Mahmoudreza Dehghan - Worked on researching basic regression algorithms and implementing them.

Additionally, it is worth noting that all team members have participated actively in editing each other's sections and the overall report as a whole.

Our github repo can be found at https://github.com/rluuy/lbt_alignment