Internship Report at Fraunhofer IGCV

Ivan Smirnov

Machine Learning für Sensorsysteme

Introduction

Name of the Intern: Ivan Smirnov
Matriculation Number: 22109653
Semester: 5th
Name and Address of the Internship Company:
Fraunhofer IGCV, Am Technologiezentrum 2, Augsburg
Start and End of Internship: 16th August 2023 - 31st January 2024

Description of the Training Company

The Fraunhofer-Gesellschaft, based in Germany, is the world's leading applied research organization. By prioritizing key technologies for the future and commercializing its findings in business and industry, it plays a major role in the innovation process. A trailblazer and trendsetter in innovative developments and research excellence, it is helping shape our society and our future. Founded in 1949, the Fraunhofer-Gesellschaft currently operates 76 institutes and research units throughout Germany. Around 30,800 employees, predominantly scientists and engineers, work with an annual research budget of roughly €3.0 billion, €2.6 billion of which is designated as contract research.

The Augsburg-based Fraunhofer Institute for Casting, Composite and Processing Technology, IGCV, conducts production-oriented research and development with a strong focus on practical use and various forms of applications and brings together its expertise in the fields of lightweight casting technology, fiber-reinforced compound materials and automated production. The expertise of about 120 scientists ranges from material science to structural mechanics, production engineering and production.

The Fraunhofer Institute for Casting, Composite and Processing Technology IGCV is strategically located amidst various technology centers, fostering a vibrant ecosystem of innovation and collaboration. This advantageous positioning, adjacent to the Augsburg University, facilitates a seamless exchange of knowledge, resources, and talent between the institute and the academic community. The proximity to these centers of excellence not only enhances the institute's access to cutting-edge research and technological advancements but also strengthens its ties with leading experts and scholars in related fields.

This job was found through Glassdoor, alongside utilizing other platforms like LinkedIn and Indeed.

Detailed Progress Report

"Samoa" Study

During the initial phase of my internship at Fraunhofer IGCV, I had the opportunity to contribute to the "Samoa" project. This project aimed at understanding the correlation between the complexity metrics of datasets and the number of images required for a person to achieve specific performance levels in machine learning models. The complexity metrics considered included parameters like the number of superpixels, JPEG complexity, image sizes, edge density, and so on.

Data Preprocessing: I analyzed both the IGCV and MVTec anomaly detection datasets. This examination was critical to ascertain their suitability for the study. I ensured the completeness and accuracy of the datasets, a process that involved checking for missing values, outliers, and ensuring that the datasets accurately represented the problem set. The datasets were cleaned to rectify any inconsistencies and errors. This step was vital in maintaining the integrity of the analysis. Finally, I integrated multiple datasets into a unified format. This task was required to ensure that the merged dataset was coherent and analytically viable.

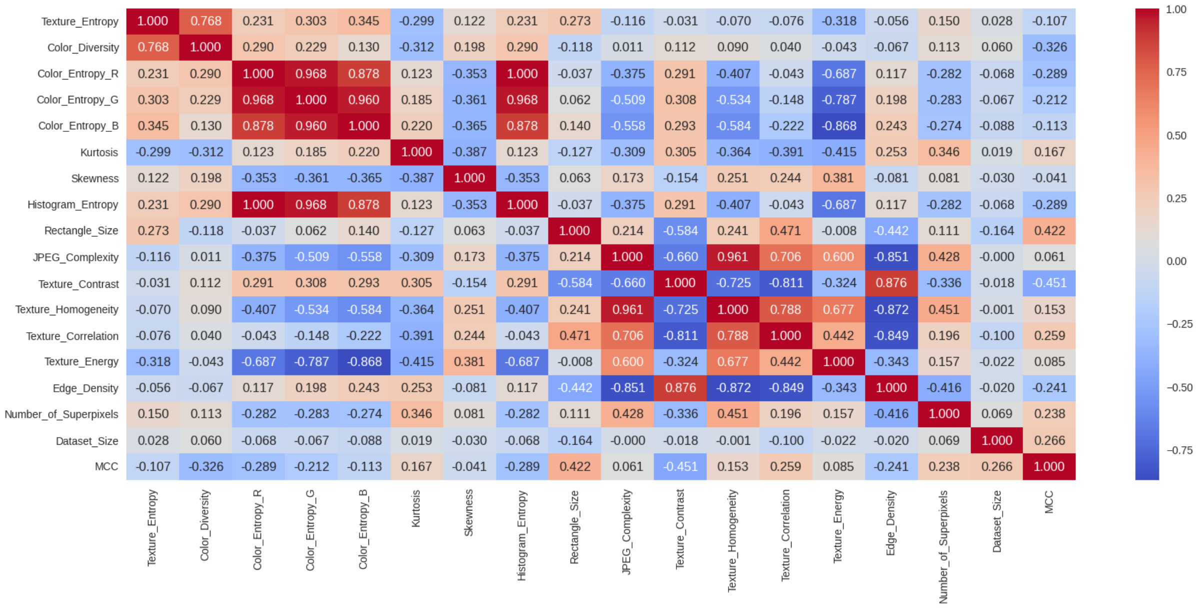

Figure 1: Correlation Matrix Between Features for the "Samoa" Project

Model Training and Data Augmentation: A significant part of my task was to train models with PyTorch to achieve fair accuracy across different datasets. This process involved selecting appropriate algorithms, tuning hyperparameters, and continuously evaluating the model's performance. I developed and applied custom data augmentation techniques. These techniques were uniquely designed to not only augment the image data but also the corresponding segmentation masks. The augmentation techniques I implemented included flips, crops, adjustments in brightness, contrast modifications, and so on. Each augmentation method was applied with a specific probability and was carefully chosen from a predefined set of types to best suit the data and objectives.

Utilization of Virtual Machines and Analysis Tools: The training processes were executed on virtual machines equipped with GPUs. This setup allowed for accelerated computation, which was essential for handling complex models and large datasets. I also used Omniboard, a sophisticated platform for analyzing training processes and metrics. This tool enabled me to track and analyze various metrics such as accuracy, loss, and other performance indicators over time.

Data Versioning: I used Data Version Control (DVC), an essential tool for managing and versioning datasets and machine learning models akin to Git for code. This approach significantly improved the project's data management, enabling us to efficiently track and update datasets, ensure the reproducibility of preprocessing steps, and facilitate seamless collaboration among team members. DVC's integration into the workflow allowed for the systematic versioning of data and models, ensuring that all changes were documented and easily accessible.

Training and Analyzing the "Samoa" Results

Following the initial phase of training models in the "Samoa" project, my focus shifted to analyzing these results to derive insights about the impact of dataset complexity on model performance. The core objective was to establish a correlation between various complexity metrics and the performance outcomes, given a specified amount of training images.

Hypothesis and Approach: Consistent with the hypothesis, it was expected that for a constant dataset size, more complex datasets would yield lower performance metrics, such as accuracy, compared to simpler datasets. I explored how reducing the training data size influenced accuracy. This was crucial in understanding model behavior in data-constrained environments.

Methodology and Analysis: To conduct this analysis, I employed traditional machine learning methods, experimenting with various algorithms to find the most suitable one. The most promising results were obtained using the Random Forest algorithm. This method was particularly effective due to its robustness and ability to handle large feature sets, which was essential for the complexity metrics analysis. Through this analysis, I was able to identify which complexity metrics had the most significant impact on the final model performance. This involved feature importance analysis within the Random Forest algorithm and through SHAP values.

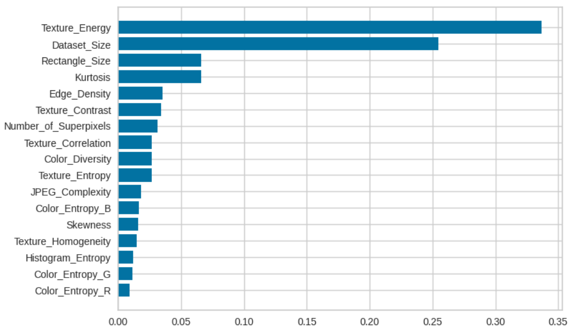

Figure 2: Random Forest Feature Importance for the ”Samoa” Project

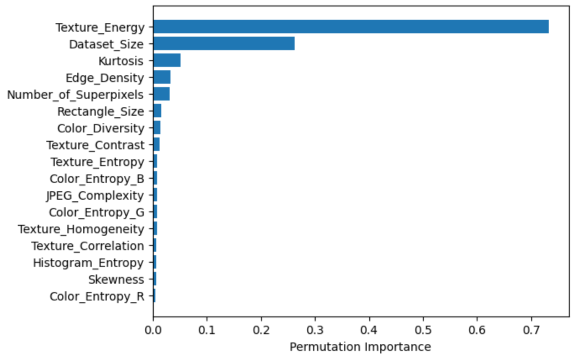

Figure 3: Permutation Feature Importance for the ”Samoa” Project

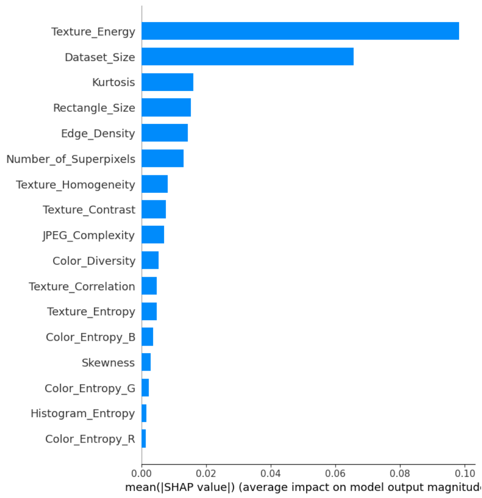

Figure 4: SHAP Feature Importance for the ”Samoa” Project

Reporting and Visualization: I compiled a detailed report documenting the findings of this analysis. The report included a thorough discussion of the methodology, results, and conclusions. To better illustrate these findings, I created various visualization plots. These included:

- Performance plots showing the relationship between dataset complexity and accuracy.
- Feature importance plots highlighting the most influential complexity metrics.

- Comparative plots to showcase model performance across different dataset sizes and complexities.

Additional Insights for the "Samoa" Project

The visualizations presented in Figures 2, 3, and 4 provide a comprehensive understanding of feature importance from different perspectives. The correlation matrix in Figure 1 highlights the relationships between various features, aiding in the identification of multicollinearity and informing feature selection processes.

The R-squared scores in Table **??** demonstrate the robustness of the Random Forest estimator against the removal of individual features. Notably, the removal of the Kurtosis feature results in a significant drop in the R-squared score, indicating its critical importance in predicting the MCC metric.

Researching and Training Deep Learning Models: "Idamo" Project

In my "Idamo" project, I was enhancing model accuracy for asphalt core layer detection. This industrial project was centered around analyzing cylinder asphalt cores to accurately predict the layers visible in their images. Initially, the project focused on processing raw images of the cores. However, as we progressed, we developed a more sophisticated approach that involved combining images from different sides of a core to create a single, comprehensive image. This method allowed us to achieve a more detailed and accurate representation of the core's structure.

The asphalt cores we worked with were sourced from various locations across Germany, Austria, and Switzerland, each presenting unique challenges. The diversity in the origin of these cores introduced a significant level of complexity due to variations in composition and condition. Some of the cores exhibited damages, leading to imperfections in the resulting images that complicated the layer detection process.

Despite these challenges, the project achieved remarkable results with simpler cores, where the layers were distinct and easily identifiable. Our models were capable of accurately predicting these straightforward layer structures, demonstrating the project's success in scenarios with clear delineation. However, the task became exponentially more difficult with cores that contained complex, ambiguous layers. These were cases where even human experts could struggle to pinpoint the exact layers with pixel-perfect precision.

The project's evolution from analyzing raw images (taking photos of a core as it is) to integrating multiple perspectives into a unified image representation was a significant leap forward, enabling us to tackle a wide range of core types. Yet, the variability in core quality and the inherent complexity of some layers underscored the challenges of applying machine learning to real-world industrial problems, where the goal is not only to develop a working model but to create a solution that can adapt to the vast diversity of real-world conditions.

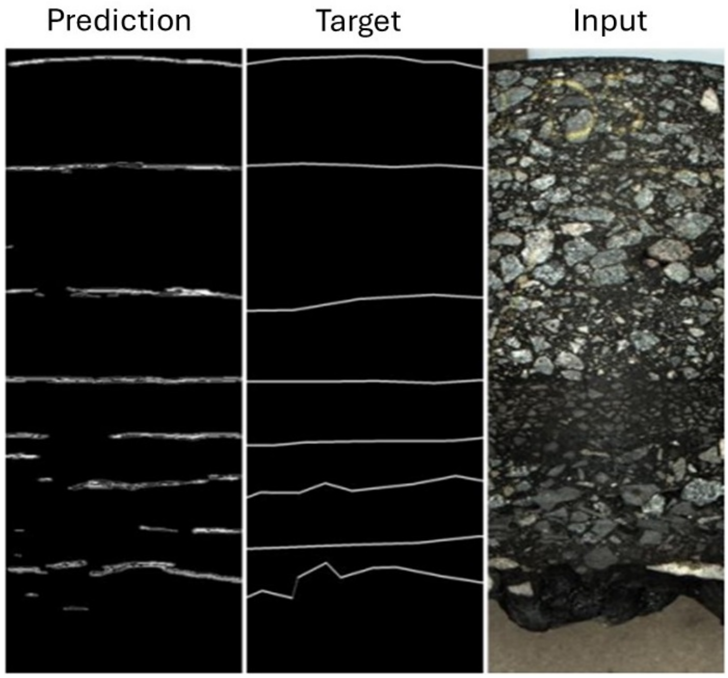

Figure 5: Process of Generating Prediction Lines in the ”Idamo” Project

Data Annotation: Within the ”Idamo” project, I was responsible for annotating new images to ensure the consistency of our dataset’s labeling, a task that required maintaining alignment with our established standards. This involved partially labeling images and carefully adjusting them to match the collective understanding of what constitutes a layer versus background, a decision that could vary significantly. Additionally, I integrated feedback from domain experts into our labeling process, an essential step for capturing nuanced details of asphalt core layers and ensuring our annotations reflected both the technical and practical realities of the field.

Model Training and Analysis: I embarked on an extensive exploration of various segmentation models, analyzing their strengths and weaknesses in the context of asphalt core layer detection. Through this comprehensive analysis, I discovered that UNet++, a state-of-the-art segmentation architecture, along with its variations, consistently outperformed other models in terms of accuracy and efficiency. The superior performance of UNet++ was attributed to its enhanced ability to capture intricate patterns and details within the images.

To further refine our model’s output, I experimented with a wide range of hyperparameters and implemented diverse data augmentation techniques. These efforts were aimed at optimizing the model’s learning process and ensuring robustness against the variability present in our dataset. The augmentation strategies, including rotations, flips, and scaling, significantly contributed to the model’s ability to generalize across different asphalt core samples, thereby enhancing its predictive accuracy.

Building on the success of the UNet++ model, I devised a novel approach that involved using its output as an additional input to a secondary model. This innovative strategy was designed to address the specific challenge of combining separate predicted masks into a single, continuous line, effectively removing any unnecessary noise from the predictions. By training this secondary model to refine and consolidate the initial predictions, we achieved a substantial improvement in the overall performance of our system. This dual-model approach not only reduced prediction errors but also produced more consistent and continuous lines.

Predicting Number of Layers and K-Means: Additionally, I created an OpenCV-based function to calculate layer numbers directly from the mask. This was necessary since the original images were too big and we had to go through smaller parts of the original image to predict the layers. This additional ability to predict the number of layers was helpful for the final layer creation: utilizing K-Means clustering to generate continuous lines from the segmented mask (K here is the predicted number of lines). As a result, all the issues with the original prediction (not continuous lines) were fixed by finding the K clusters of segmentation masks and combining separate intermittent parts of the layer into a single one.

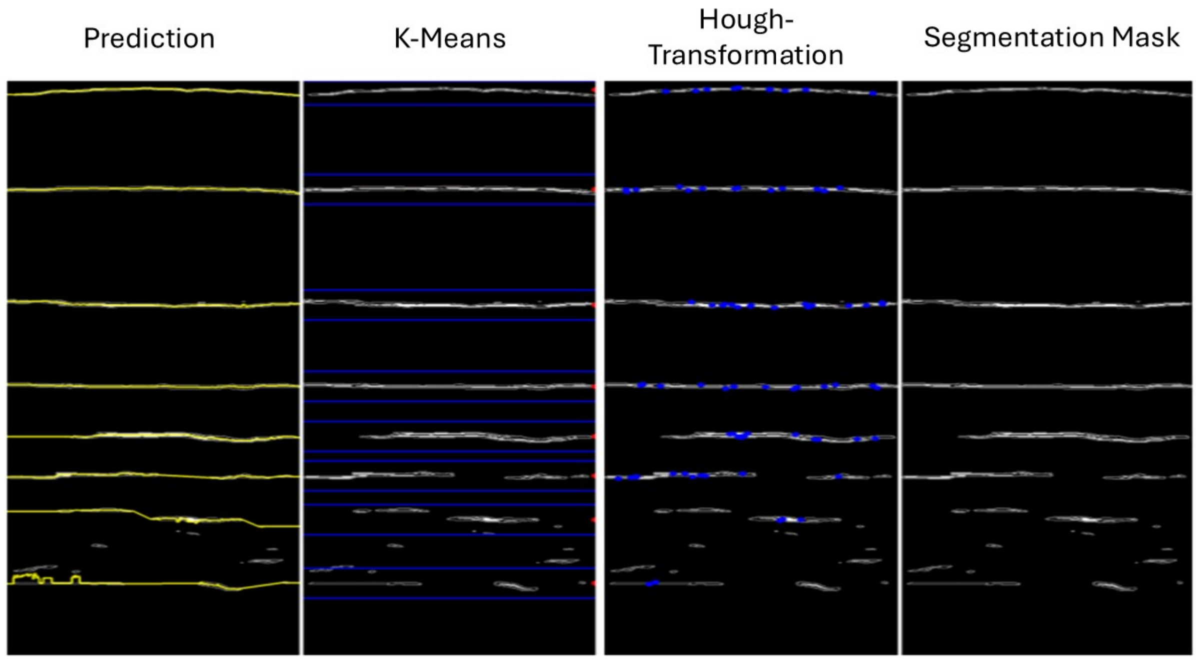

Figure 6: Overview of the Asphalt Core Data in the "Idamo" Project

Avoid Stones: Further, I developed an algorithm using the A* search strategy to draw final lines that navigate around obstacles, such as stones. This was achieved by employing OpenCV to create a stone mask using Adaptive Thresholding, then applying A* to find the shortest path avoiding the stones. This algorithm played a crucial role in enhancing the precision of the segmentation task by ensuring that the predictive lines accurately represented the asphalt layers without intersecting any detected stones.

Additional Insights: Figure 6 illustrates the diversity and complexity of the asphalt core images used in the "Idamo" project. The variation in core conditions and compositions from different regions introduces significant challenges in accurately detecting and predicting layer structures.

Figure 5 depicts the innovative method developed for generating continuous prediction lines. This process involves integrating multiple segmentation masks and applying advanced algorithms like K-Means clustering and the A* search strategy to ensure precision and continuity in the detected layers, effectively navigating around obstacles such as stones.

Code Refactoring for "Idamo" Project: My responsibilities also included code refactoring for the "Idamo" project, which involved:

- **Code Optimization**: Refactoring existing code for better efficiency and readability.
- **Documentation**: Adding comprehensive comments to enhance understanding and maintainability.
- **File Organization**: Restructuring file locations for more logical and intuitive access.

Final Assessment of the Internship and the Training Company

During my internship at Fraunhofer IGCV, I worked on expanding my expertise in both deep learning, using PyTorch, and classical machine learning. This aspect of my work was not previously covered in my coursework.

In terms of project development for the "Idamo" project, the rate of data acquisition was quite slow. Initially, my work involved researching new methods for layer detection, leading to the realization that acquiring high-quality data was crucial for progress.

Additionally, I applied traditional algorithms, such as the A* algorithm, to create solutions for drawing lines that avoid obstacles like stones. This work emphasized the importance of leveraging a variety of methods in AI and machine learning projects.

This internship, while immensely educational, also presented substantial challenges, particularly due to imperfections in the data and the speculative nature of some research components, such as the investigation into complexity metrics, which carried no guarantee of success. These hurdles underscored the difficulties of working with imperfect datasets and trying untested methods, where outcomes are uncertain. Despite these obstacles, we made considerable progress in the "Idamo" project. However, due to these complexities and the slow pace of data acquisition, the project remained incomplete by the end of my internship.

Overall, the internship enhanced my skills in multiple areas of AI and machine learning, providing a comprehensive learning experience.