THE HAGUE
UNIVERSITY OF
APPLIED SCIENCES

Research Plan

Automation of an image processing software
for images of nanoparticles using Machine Learning models

Klara Baumeister 19029454
Yoran de Vos 17049784
Oscar den Buurman 17006465

Problem owner: Tomas Storck
Minor Applied Data Science

Research Plan

Nano Group

Introduction

In coherence with our minor Applied Data Science at The Hague University of Applied Sciences, our group was introduced to VSParticle's research objective as our course project. VSParticle is a company that derived from TU Delft's nanotech group. They work on nanoparticles and, among others, analyze nanoparticle images to calculate the particle's sizes. To do this, VSParticle has built an image processing program that results in an edited image which enables the software to calculate the particle's sizes easily. Part of this program is the thresholding step, where greyscale images are converted into bitmaps using thresholding algorithms and user input. In combination with our minor, our goal is to automate this step using a Machine Learning model which predicts the best algorithm to use. Our research question reads as follows:

How can a Machine Learning model, that predicts the optimal thresholding algorithm, assist VSParticle to analyze nanoparticle images?

To answer our research question in detail, we are going to focus on the following four sub-questions:

1. How does the given data need to be restructured in order to be useful for the model?
2. What features of the dataset should be selected for the model?
3. What type of model do we need and how is it structured?
4. How can the predictions of the model, graded by VSParticle's user, be employed to improve the model over time?

In order to achieve our goal, we have received a json file from VSParticle containing IDs of each run (a run being an image being processed once), parameters of the run, metadata of the image, resulting images of each step, and scores. The scores are our point of focus: they contain computer generated values that evaluate how well each step of the program worked, as well as a user score, given manually by the user after each run. This user score is crucial for our approach: We plan on predicting the user score for each available thresholding algorithm to determine which option gives the best results. In short, the thresholding algorithm that leads to the best user score will be selected.

To make working in a group as efficient as possible, we are going to use Scrum. This allows us to work in a flexible, dynamic and focused way. Having sprint periods of two weeks, including daily standups and retrospectives at the end of each sprint, allows us to work in small, efficient steps and to have clear communication and a good overview of the project. Our planned schedule is as follows:

Research Plan

Nano Group

Week	Sprint	Sub-question	Plan
1	1	Sub-question 1	Prepare and plan project
2			Prepare data
3	2		Visualize data
4			
5	3	Sub-question 2	Determine approach
6			Create first simple models
7	4		Experiment with different ML models and features
8			
9	5	Sub-question 3	Research on handling imbalanced data
10			Deep Learning model on Yen
11	6		Revisiting json file + Decision Tree + Logistic Reg.
12			Experiment: Find best model combination
13	7	Sub-question 4	**Present final model to problem owner**
14			(Maybe: add more data, briefly look into images)
15	8		Research paper
16			Research paper
17	9		Research paper
18			**Finalize research paper + present it**

1. How does the given data need to be restructured in order to be useful for the model?

Aims

The data that is given by VSParticle must be structured in a way so that it can be used to create a model. The aim of the research is to:

- Provide information on the current structure.
- The most optimal structure for the model.
- Methods to transform the data into the most optimal structure.

Methodology

The research will be done via articles, experimenting in pytorch and lectures. When a method is introduced it will be applied to our dataset and tested to see what kind of impact it had. This will be a cycle until the best possible solution is found.

Expected outcomes

The found methods to restructure the data will transform the data in a way that it positively affects the accuracy of the model.

Timetable

This research will be done throughout the whole project as the model will change over time.

2. What features of the dataset should be selected for the model?

Aims

The model should make use of the most helpful features in order to be as successful and accurate as possible. The aim of the research is to:

- Find the features closest related to the thresholding step
- Determine how many features lead to the best results
- Find the most meaningful combination of multiple features

Methodology

The research will be done through experimentation in Python and data analyzation. Applying the knowledge that we gained in lectures as well as calculating coefficients and accuracy scores for possible models and feature combinations will hopefully lead to meaningful results which help us determine the best possible features to select. Literature research on how to analyze and select features could therefore also be helpful for us.
Use script that shows all possible feature combination's correlation

Expected outcomes

We expect to determine two to six features that are most closely related to the thresholding step and, applied in combination, lead to an accurate model. Furthermore, we expect the features to be scores drawn from the thresholding step, as these are directly linked to the step we want to predict.

Timetable

This research will be done before and simultaneously as we create the model, in order to have a good sense of direction of what we should focus on and to be able to reiterate over our feature choices.

3. What type of model do we need and how is it structured?

Aims

For this project, we aim to find and employ the most accurately fitting model. Since the model type determines how it operates in general, it also determines whether or not it is successful in combination with our data. The aim of this research therefore is to:

- Discover what type of model suits the data best (e.g. Linear Regression, Logistic Regression, Classification, Neural Network, …)
- Find out what kind of structure leads to the best results

Methodology

The research will be done through combining theoretical knowledge we gain in lectures, literature research and experimentation in Python. The lecture knowledge and literature research will help us select possible model types and applying these model types on our dataset and experimenting with them will help us detect the most useful one.

Expected outcomes

We expect our research to result in one model type that hopefully clearly fits our data and is able to make reliable and accurate predictions; or multiple possible model types of which one shows clear benefits over the others. After this, we expect to determine a structure with which the model type works best.

Timetable

This research will be done before we create the final model, as the final model will make use of our research and apply the model type we selected.

4. How can the predictions of the model, graded by VSParticle's user, be employed to improve the model over time?

Aims

The model should be able to continue changing and improving as more data is collected over time. This can be achieved by collecting data from each new run, including the user scores given manually at the end. The aim of the research is to:

- Find a way for VSParticle to feed newly generated data to the model easily
- Making the model "future-proof"
- Determine over time changing accuracy and usefulness of the model

Methodology

The research will be done through literature. Finding other projects that improve over time, analyzing and evaluating their method will help us gain insight into what direction of focus will be most helpful. Since this project is part of our minor, we are limited in time. Therefore, we can only prepare our model for more data to come, and will have to leave the implementation to VSParticle, when they actually do collect more data.

Expected outcomes

We expect to find a way to prepare our model for newly added features and data, therefore making sure it is able to improve in the future.

Timetable

This research will be done as one of the last steps of our model creation, as it is not essential to the model itself and focuses on future improvement. It can only be done after the initial model creation as we need to have a working, reliable model to start with before we implement this feature.