Using Explainable AI to Analyse and Improve the Behaviour of an Image Segmentation Model Trained on Synthetic Data

Project Plan

Patrick von Velasco

October 17, 2020

Motivation

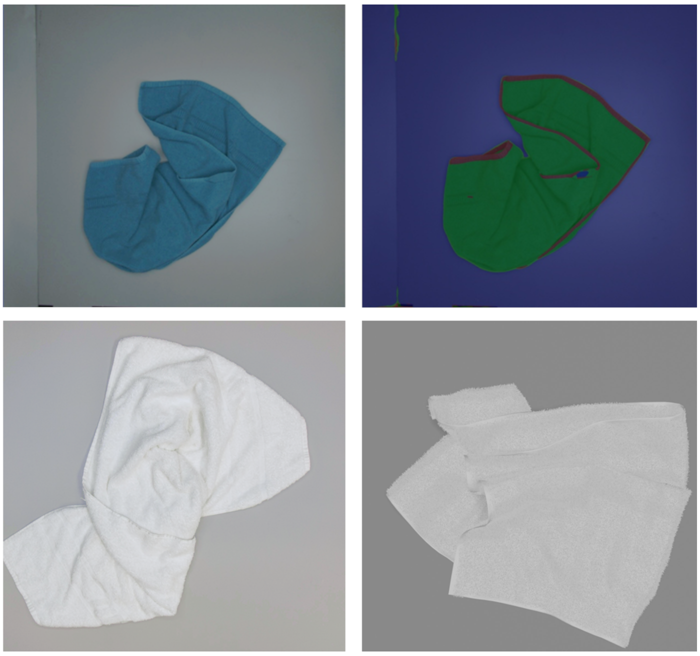
In recent years, convolutional neural networks (CNNs) have become the main method used to solve image classification and visual object detection tasks. A major drawback of using CNN architecture is that they usually require a large amount of labeled training images to train, which is why the labor intensive task of image annotation is a significant hurdle for many applications. One method to circumvent this is by generating synthetic training data and using the model – trained on synthetic data – for the real task (transfer learning). Often when this approach is used, the performance of the model when evaluating it on real data is significantly worse. This effect is known in research as the “reality gap”(1).

1 Project Description

Sewts GmbH[1] is a TUM-based startup focused on delivering computer vision and robotics solutions for textile handling. One big area of research at the company is identifying different states of a textile - specifically identifying the seam areas - from image data.// For this task, the company currently uses CNN-based image segmentation algorithms. Since the task at hand requires detailed annotation of each area of the image, generating ground-truth data is a very labor-intensive process. Therefore physical finite element method (FEM) simulation software is used to render synthetic textile images to increase the amount of available training data (See Figure 1).

As it currently exists, the model does not translate well to real data when trained on synthetic images. This "reality gap" is a common phenomenon encountered during machine learning tasks. Since the currently used CNN model and training process is largely a "black-box" and the parameters governing it are determined more by educated guesses than by scientific process, no procedures for investigating the model's stability and reliability exist at the company.

Figure 1: Top left to bottom right: Example image of towel on table, seam segmentation map of the towel (model output), real and simulated training image



Therefore **this thesis aims at using "explainable AI" concepts and methods to investigate the model in use and it's reality gap, comparing the model's behavior and stability when trained on real vs synthetic data, as well as using the insights from this process to research and compare methods of augmenting the synthetic data.**

[1] https://www.sewts.de/

In short, we want to answer the question **“why does the model, when trained on synthetic data, not translate to real data and what can we do to improve that behavior?”**

2 Methodology

In order to answer the above research question, the following steps are proposed:

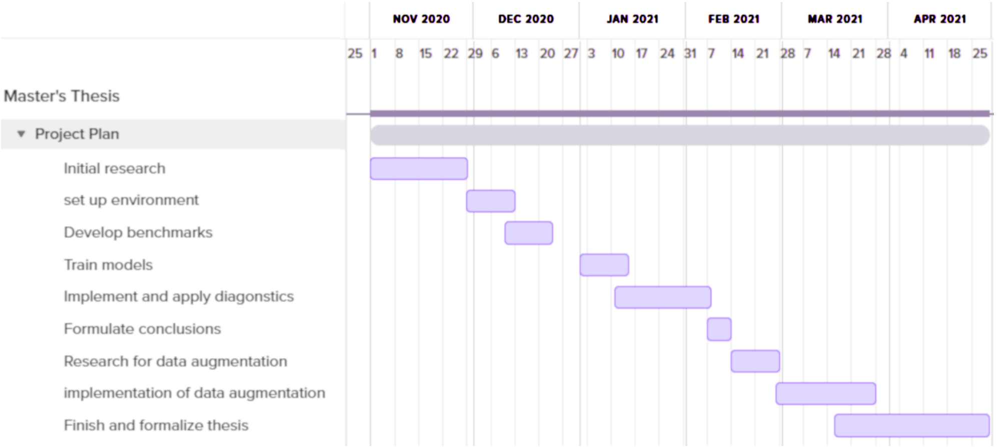
1. **Research state of the art diagnostic methods for interpret ability of image segmentation/classification CNN.** This includes - but is not limited to – layer visualizations, saliency maps(2), class activation mapping(3) and SHAP-scores(4). The researched methods should be compared and their suitability for the given task should be discussed.
2. **Acquire the neural network model and set up a testing environment.**The model (with fixed architecture and hyperparameters) will be treated as a fixed black-box to be investigated. For this, it might be necessary to port it to different frameworks (Keras/Tensorflow, Pytorch, etc.) depending on the methods of choice found in point 1. and the availability of suitable toolboxes. This also includes setting up a pipeline for the benchmark process.
3. **Develop a benchmark and dataset.** This includes finding real and synthetic images that are representative and comparable, separating them into testing, training and validation sets, and choosing metrics for the comparison. If necessary, additional data might need to be labeled.
4. **Train and compare models trained and validated on different combinations of real and synthetic data.**
5. **Apply the diagnostic methods chosen in point 1 to the models.** Investigate if any meaningful differences can be identified. Try to draw conclusions about the feature importance and behavior of the seam segmentation task.
6. **Depending on the results and findings, propose and/or implement methods to improve the model’s response when trained purely on synthetic data.** This point is kept open deliberately, as it is dependent on the process and findings of the earlier parts of the thesis. At the least, several methods to augment the synthetic data will be discussed and their application to the seam segmentation task will be compared. These might include, but are not limited to:
    - Adding real-life camera distortions (chromatic aberration, depth-of-field) to the synthetic images
    - Using style transfer methods to make the synthetic images more like the real ones
    - Adjusting rendering parameters during the training process(5). This might include domain randomization techniques(1)

3 Time Planning

Based on the steps above, the following rough time estimates will be given:

1. Conduct initial literature research	21-28 days
2. Set up testing environment and conditions	7-14 days
3. Develop benchmark and dataset	7-14 days
4. Train models	7-14 days
5. Investigate the models	
a) Implement diagnostic methods	14-28 days
b) Apply methods to the trained models	7-14 days
6. Conduct research towards improving performance on synthetic data	
a) Draw and write up conclusions	7 days
b) Research methods for augmenting synthetic data	7-14 days
c) Implement and evaluate data augmentation, largely dependant on the amount of time left.	14-28 days
7. Finish and formalize the thesis and presentation	45 days

Figure 2: Gannt Chart with rough time estimate for the work packages



4 Risk Analysis

Since a large part of the thesis is aimed at a relatively open investigative question, there is a chance that the methods used will not turn up any insightful results for the specific research question posed. Still, the work that went into comparing and adapting CNN tnterpretability methods and frameworks for the given object segmentation task can be a meaningful contribution either way. The data augmentation part is currently planned as an addendum to the main work of the thesis (max. 25%), but this might change depending on both the timing and results of the previous work.

5 Changes in the Project Plan

All occurring changes of this project plan will be recorded in this table in order to track deviations or changes of the planned goal.

#	Date	Change	Reason
1			
2			
3			
4			
5			

References

[1] J. Tremblay et al., “Training deep networks with synthetic data: Bridging the reality gap by domain randomization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2018.

[2] K. Simonyan et al., “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *Workshop at the International Conference on Learning Representations (ICLR)*, 2013.

[3] B. Zhou et al., “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 2017.

[5] H. S. Harkirat et al., “Autosimulate: (quickly) learning synthetic data generation,” in *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020.