**Capstone Project** **Document**

Sang Don Lee

Cohort Part Time 23rdJune2020

3/12/2020

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.

Diagram

Description automatically generated

# Problem statement

* What is the problem or the opportunity that the project is investigating?
  + How do the textile/fabric manufacturers predict defects on the textile during its manufacturing?
* Why is this problem valuable to address?
  + Because they rely on human resources to detect textile defects and its accuracy is around 70% or less, so it is critical for them to improve the detect rates for their better-Quality Control process that leverages the Quality Assurance.
* What is the current state (e.g. unsatisfied customers, lost revenue)?
  + They are losing their reputation of their product quality by supplying defective textile and the cost of returning them goes up, impacting on customer satisfaction which ends up losing their business and market share.
* What is the desired state?
  + They wanted to improve the detection rate over 90% with minimal of false detection which is not critical, however they want to keep the non detecting rate on the defective textile.
* Has this problem been addressed by other research projects? What were the outcomes?
  + Yes, a research paper published below using the same dataset:
  + <https://content.sciendo.com/view/journals/aut/19/4/article-p363.xml>

# Industry/ domain

* What is the industry/ domain?
  + Manufacturing
* What is the current state of this industry? (e.g. challenges from startups)
  + Challenges on improving the QC and raising cost for returning defective goods
* What is the overall industry value-chain?
  + Raw material -> Process -> Quality management -> Supply to customers.
* What are the key concepts in the industry?
  + Improving quality of product, minimising costs, maximise the market share
* Is the project relevant to other industries?
  + Yes, other manufacturing industry as a part of Smart factory start up

# Stakeholders

* Who are the stakeholders? (be as specific as possible)
  + Manufacturing owner, QC Managers/Staff, and procurement
* Why do they care about this problem?
  + Supplying defective textile/fabric impacts on multiple factors such as raising costs of returning goods, losing reputation of the goods, unhappy customers and losing market share
* What are the stakeholders’ expectations?
  + By implementing autonomous QC process to reduce the manual QC errors

# Business question

* What is the main business question that needs to be answered?
  + We want to initiate Smart Factory transform for a long term.
  + We have around 70% of textile defect detection rate based on skilled QC Staff.
  + For a start, how can we implement automated textile defect detection and improve its rate?
* What is the business value of answering this question? (quantify value and make necessary assumptions)
  + Not an exact quantify value however addressing the percentage in comparison to the exiting workflow.
  + Increasing +29% from existing defect detection rate (less than 70%)
  + This should reduce returning cost of textile
  + Intangible value of customer satisfaction
  + Reutilising exiting QC staff to allocate them for more creative work
* What is the required accuracy? What are the implications of false positives or false negatives?
  + Having FP is relatively less critical as the model predicts it as defective, however FN is very critical when the model is not able to detect the defective textile. The model accuracy/recall are performing around 99% on both metrics.

# Data question

* What is the data question that needs to be answered?
  + Model can predict the defect images?
* What is the data required to answer the question?
  + Non-defect samples
  + Defect samples and defect pattern (Masks)

# Data

* Where was the data sourced?
  + <https://www.aitex.es/afid/> : AFID: a public fabric image database for defect detection.
* What is the volume and attributes of the data?
  + 140 Non-defect images
  + 107 Defect images
  + 107 Mask images
* How reliable is the data?
  + It is very reliable as it was from a research company collected from real samples.
* What is the quality of the raw data?
  + Very good, but the image size is very interesting (4096 x 256)
* How was this data generated?
  + As described above in “where was the data sourced?”
* Is this data available on an ongoing basis?
  + Yes

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
  + Image pre-processing
    - Resizing
    - Normalisation
    - Augmentation
  + Balancing the data
* What are the highlights of the Exploratory Data Analysis (EDA)?
  + Image size
  + 12 different defect types
  + Defect and Mask mapping
* Is the pipeline reusable? (for example, to process future data?)
  + Yes
* What are the intermediary data structures used (if any)?
  + trained with only resized images without EDA to observe if the model is viable and its performance

## Modelling

* What are the main features used?
  + NA as this is image dataset
* Did you find any interesting interactions between features?
  + NA
* Is there a subset of features that would get a significant portion of your final performance? Which features?
  + NA
* How did you select features?
  + NA
* What feature engineering techniques used?
  + Image pre-processing
    - Resizing
    - Normalisation
    - Augmentation
  + Balancing the data
* What are the models used?
  + CNN Sequential
  + CNN Inception
* How long does it take to train your model?
  + 2.5 hours with GPU enabled using a Kaggle notebook
* What are the tools used? (cloud platform, for example)
  + Kaggle notebook
  + Colab notebook
* What are the model performance metrics?

A picture containing treemap chart

Description automatically generated

* Which model was selected?
  + CNN Inception

## Outcomes

* What are the main findings and conclusions of the data science process?
  + Inception CNN model is very highly performing the prediction
  + Image pre-processing and relative feature engineering techniques are very important and time consuming

## Implementation

* What are the considerations for implementing the model in production?
  + The manufacturers are able to implement the model in their production line relatively easily, however they have to invest some funds to implement the model in terms of E2E implementation \*\* this explains in the section of “End-to-end solution”

# Data answer

* Was the data question answered satisfactorily?
  + Yes
* What is the confidence level in the data answer?
  + 99% of Recall represents very high confidence level of the model performance. Also proves the model predicts very small number on FN (False Negative) which is a critical factor for this model.

# Business answer

* Was the business question answered satisfactorily?
  + Yes, with very high Recall and very small on FN which can implement on their autonomous QC process
* What is the confidence level in the business answer?
  + Same as above.

# Response to stakeholders

* What are the overall message and recommendations to the stakeholders?
  + The model can be implemented to their production line in accordance with other requirements

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?
  + Addressed very high level E2E solution
  + Requires cameras or cams capturing images
  + Collecting images and ship them to where the model runs
  + The environment for the model can be either on public clouds or on premise
  + Providing the User interface to visualise the images, provide statistics

Graphical user interface, application

Description automatically generated

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
  + <https://github.com/sangdonlee/capstone>
* What are the resources used in the project? (libraries, algorithms, etc)
  + OpenCV, glob, pyplot etc
  + Tensorflow for image processing/import/algorithm
  + Algorithm referenced: [www.medium.com](http://www.medium.com)
    - https://medium.com/%E9%9B%9E%E9%9B%9E%E8%88%87%E5%85%94%E5%85%94%E7%9A%84%E5%B7%A5%E7%A8%8B%E4%B8%96%E7%95%8C/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-ml-note-cnn%E6%BC%94%E5%8C%96%E5%8F%B2-alexnet-vgg-inception-resnet-keras-coding-668f74879306
  + Stack overflow for code tips
  + Video training source: Datacamp and Youtube ML/Deep Learning/CNN