Quantity beats quality for semantic segmentation of corrosion in images

Will Nash∗  Tom Drummond†  Nick Birbilis‡

July 10, 2018

Abstract

Dataset creation is typically one of the first steps when applying Artificial Intelligence methods to a new task; and the real world performance of models hinges on the quality and quantity of data available. Producing an image dataset for semantic segmentation is resource intensive, particularly for specialist subjects where class segmentation is not able to be effectively farmed out. The benefit of producing a large, but poorly labelled, dataset versus a small, expertly segmented dataset for semantic segmentation is an open question. Here we show that a large, noisy dataset outperforms a small, expertly segmented dataset for training a Fully Convolutional Network model for semantic segmentation of corrosion in images. A large dataset of 250 images with segmentations labelled by undergraduates and a second dataset of just 10 images, with segmentations labelled by subject matter experts were produced. The mean Intersection over Union and micro F-score metrics were compared after training for 50,000 epochs. This work is illustrative for researchers setting out to develop deep learning models for detection and location of specialist features.

Keywords—Machine Learning, Corrosion, Semantic Segmentation, Datasets, Fully Convolutional Network

1 Introduction

Corrosion is a difficult subject to detect compared to other common subjects such as the human face that have distinct features: two eyes, a nose and a mouth; corrosion shares limited characteristics in colour range and texture - and the appearance of corrosion is confused on both counts, with shadows, boulders, bricks, safety vests all presenting false positives for Deep Convolutional Neural Network (D-CNN) models. Furthermore, the boundary between corroded and uncorroded areas in images is often undefined, due to image compression artefacts and focal range (depth of field).

To produce a useful detector requires semantic segmentation of images rather than simply image classification, i.e., per pixel labelling rather than image level labelling. To achieve this a Fully Convolutional Network (FCN) (Long et al. 2015) was employed using strongly supervised training; requiring a dataset of images labelled as densely as the desired output. Ideally, the model can incorporate expert level decision making regarding the severity of corrosion detected. Creating a dataset for Deep Learning of corrosion segmentation requires subject knowledge from labellers to avoid mislabelling of ground truths. The current research is investigating the efficacy of D-CNN for automated detection of corrosion in rapid infrastructure inspections. Within the present paper the balance between a large dataset with poorly labelled ground-truths and a small dataset with expertly annotated segmentations is investigated.

1.1 Related Work

Impressive progress in D-CNN for image classification tasks have been driven in large part by the availability of massive labelled datasets such as ImageNet (Jia Deng et al. 2009). Within the field of semantic segmentation, where each pixel is assigned a class prediction by the model, the largest
public datasets include PASCAL VOC (Everingham et al. 2010) and MS-COCO (Lin et al. 2014), labelled via a supervised annotation event for the former, and leveraging the Amazon Mechanical Turk (with a specialised user interface) for the latter. These large datasets have an advantage in that the subject matter is easily recognizable by the general public, MS-COCO claims the objects are easily recognizable by a 4-year old. When it comes to specialised subject matter labelling the dataset becomes much more challenging - for example the BRATS dataset of brain tumours (Menze et al. 2016) was produced from just 65 MRI scans annotated by seven expert radiographers, with each segmentation taking approximately 60 minutes.

The difficulty in producing semantic segmentation datasets is recognised and tackled by many researchers using novel approaches to automate dataset creation. The simple-to-complex approach (Wei et al. 2015) uses saliency mapping from image classification to generate a dataset for training a simple semantic segmentation model, that then produces a dataset for a more complex semantic segmentation model. (Bearman et al. 2016) use point clicks on subject objects as a semi-automated method to generate rough semantic segmentation maps. Otherwise fine-tuning of pre-trained networks on specialist subjects has proven successful (Milan et al. 2017), especially where the specialist dataset shares common lower level features with the pre-trained dataset.

Mislabelled data was found to negatively impact classification performance more than if it was excluded from the dataset by (Reale et al. 2016) - with just 10% mislabelled data roughly equivalent to halving the dataset size. However, (Rolnick et al. 2017) found that larger model architectures are able to deal with significant label noise provided it is not adversarial, achieving good performance with less than 1% better than chance on classification labels - although a maximum useful size of dataset was found, beyond which performance improvements were marginal. Intuitively, larger neural network architectures are able to learn more representations of features, which provides an increase in robustness to noisy training signals. The relationship between dataset size and classification accuracy has also been investigated by (Cho et al. 2015) who found that increasing dataset size increases accuracy, and formulated a method to estimate the dataset size required to achieve target accuracy for a specific task by training on varying dataset size.

Incorrectly labelled semantic segmentation features can be considered adversarial if they are detrimental to learning, i.e., they confuse the model by providing feature level false positives and negatives in the training set. Whereas, non-adversarial noise comprises edge cases where the boundary is not clear, these can be overcome by the sheer number of correctly labelled pixels. To the authors’ knowledge no previous work has been undertaken to evaluate the impact of label noise on deep learning semantic segmentation models.

2 Methodology

Two competing Datasets (DS) were produced to test the efficacy of training with a large ‘imperfect’ DS against the performance of training with a small ‘perfect’ DS:

- DS-A is used to denote the larger data set of 250 images, imperfectly labelled for two classes: corrosion and background. An 80 / 20 split was used for training and validation sets respectively.
- DS-B refers to the smaller data set of ten images, expertly segmented for corrosion in four classes: minor, moderate and severe corrosion, plus background. Training was undertaken on five images with two images reserved for validation.
- Three ‘Assessment’ images were reserved and expertly segmented for the performance comparison of the two models. Figure [1] presents an example assessment image and the associated labelled segmentations for the two models.

The FCN used herein was based on the modified VGG-16 architecture (Simonyan & Zisserman 2014). Python code was adapted from the TensorFlow implementation of (Teichmann et al. 2016), which was successfully used for semantic segmentation of road scenes from the KITTI autonomous driving dataset (Fritsch et al. 2013). The FCN was trained using transfer learning for 50,000 iterations on an NVIDIA GTX 1080Ti with the following hyper-parameters: Optimizer: Adam, Learning Rate: $10^{-5}$, Batch Size: 1, Loss: Cross Entropy.
Performance of the two models was assessed by comparing the confusion matrices across classes; and calculating the mean Intersection over Union (mIoU) and micro F-scores for the three assessment images. To provide a fair comparison the DS-B metrics were computed for ‘all corrosion’, where confusion between the corrosion classes was ignored. The ‘background’ class was excluded from the metrics because it vastly outnumbers the corrosion pixels.

\[
m\text{IoU} = \frac{1}{N \sum_{2}^{N} \frac{TP}{(TP + FP + FN)}}
\]

\[
\text{micro F-score} = \frac{\sum_{2}^{N} 2TP}{\sum_{2}^{N} (2TP + FP + FN)}
\]

Equations 1 and 2 present the formula for computing the mIoU and micro F-score respectively; where: \(TP = \text{True Positives}, FP = \text{False Positives}, \) and \(FN = \text{False Negatives}\).
3 Results

The three assessment images and their respective model predictions are presented in Figure 2.

Figure 2: Output predictions of models. (a), (b), and (c): assessment images; (d), (e), and (f): DS-A prediction; green = corrosion; (g), (h), and (i): DS-B prediction; green = minor corrosion; red = moderate corrosion; and blue = severe corrosion.
Figure 3 presents the accuracy of the output predictions from the models compared to the expertly segmented labels. The corrosion prediction accuracy was generated by multiplying the background class segmentation in red by the background class prediction in cyan.

![Image](image-url)

**a: DS-A Image 1 Accuracy**  
**b: DS-A Image 2 Accuracy**  
**c: DS-A Image 3 Accuracy**  
**d: DS-B Image 1 Accuracy**  
**e: DS-B Image 2 Accuracy**  
**f: DS-B Image 3 Accuracy**

Figure 3: **Comparison of DS-A and DS-B models accuracy.** (a), (b), and (c) show the accuracy of the DS-A model; (d), (e), and (f) show the accuracy of the DS-B model. White = True Positive, black = True Negative, red = False Negative, cyan = False Positive

The confusion matrices for the DS-A and DS-B model outputs of the three assessment images are provided in Tables 1 and 2 respectively. For the DS-B model the per class metrics and combined ‘all corrosion’ class metrics (in bold) are presented.

**Table 1: DS-A Confusion Matrix:** Metrics to assess accuracy performance of the DS-A model, including mean Intersection over Union (mIoU) and F-score

| Image | Prediction | Label | mIoU  | F-score |
|-------|------------|-------|-------|---------|
|       | background | 4,768,771 | 251,161 | 0.43    | 0.60    |
|       | corrosion  | 62,883  | 237,235 | 0.32    | 0.56    |
|       | background | 4,654,725 | 295,996 | 0.53    | 0.69    |
|       | corrosion  | 18,857  | 350,422 | 0.22    | 0.36    |
|       | background | 4,603,629 | 431,649 | 0.22    | 0.36    |
|       | corrosion  | 124,488 | 160,234 |         |         |
|       | Average    | 0.39    | 0.55   |         |         |

An example output and the accuracy of the multi-class model trained on DS-B is presented below (Figure 4). Again, the accuracy images were generated by multiplying the label segmentation in red with the prediction segmentation in cyan.
Table 2: **DS-B Confusion Matrix**: Compilation of metrics to assess accuracy performance of the DS-B model, including mean Intersection over Union (mIoU) and F-score. The multiple classes of corrosion (minimum, moderate and severe) were combined into 'all corrosion' (bolded) to provide a fair comparison with the DS-A model.

| Image | Prediction | Label | mIoU | F-score |
|-------|------------|-------|------|---------|
|       | background | min.  | mod. | dev.    | all corr’n |
| 1     | background | 4,818,964 | 174,523 | 32,033 | 7,230 | 213,786 |
|       | min.       | 48,021 | 25,076 | 22,784 | 9,617 | 0.09 | 0.16 |
|       | mod.       | 33,451 | 735 | 11,077 | 9,668 | - | 0.07 | 0.13 |
|       | dev.       | 294,408 | 1,252 | 50,288 | 45,873 | 0.11 | 0.20 |
|       | all corr’n | 379,880 | - | - | - | 176,370 | 0.23 | 0.17 |
| 2     | background | 4,839,149 | 9,027 | 74,048 | 43,601 | 126,676 |
|       | min.       | 16,248 | 456 | 28,170 | 6,561 | - | 0.01 | 0.01 |
|       | mod.       | 73,221 | 1,494 | 56,621 | 7,916 | - | 0.20 | 0.33 |
|       | dev.       | 95,918 | 1,588 | 46,699 | 19,283 | 0.09 | 0.16 |
|       | all corr’n | 185,387 | - | - | - | 168,788 | 0.35 | 0.26 |
| 3     | background | 5,017,894 | 31,734 | 20,993 | 6,130 | 58,857 |
|       | min.       | 82,131 | 47,002 | 20,365 | 5,604 | - | 0.25 | 0.40 |
|       | mod.       | 76,959 | 2,715 | 4,076 | 2,933 | - | 0.03 | 0.06 |
|       | dev.       | 236 | 224 | 624 | 380 | 0.02 | 0.05 |
|       | all corr’n | 159,326 | - | - | - | 83,923 | 0.28 | 0.27 |

Average 0.29 0.23

Figure 4: **Performance of DS-B multi-class model.** (a) expert segmentation: white = minor corrosion; green = moderate corrosion; black = severe corrosion; and red = background. (b) prediction: green = minor corrosion; red = moderate corrosion; and blue = severe corrosion. (c) to (f) individual class accuracy: black = True Positive, white = True Negative, cyan = False Negative, red = False Positive.
4 Discussion

The first observation to be made is that neither model achieves human level accuracy for detection of corrosion. This demonstrates the difficulty of corrosion as a subject of image detection. Several methods are available to improve the accuracy of detection, such as multi-task learning (Dharmasiri et al. 2017) to help the model contextualize the information.

Generally, the micro F-scores indicate that training with a large, noisy dataset is better for semantic segmentation than training on a small expertly labelled dataset. Implicitly, the noise from the DS-A dataset is overcome by having more data for training, whereas limited data available from DS-B is insufficient to train to accurate levels.

The performance drops considerably for the DS-B model on Image 3. While similar training images are included in DS-A, DS-B has no such images indicating that the DS-B strategy suffers from over-fitting. The issue of over-fitting on small datasets is well recognised and documented in (Taigman et al. 2014) for faces, (Lu et al. 2016) for medical imaging, and an excellent discussion of the issue of over-fitting can be found in (Babyak 2004). Recent work by (Shwartz-Ziv & Tishby 2017) indicates that smaller datasets are prone to over-fitting as the model compresses and discards extraneous information in the network. Although the expert labelling provides more information, DS-B provides a total of only 27 million pixels, and these pixels do not provide independent information, therefore over-fitting with 134 million parameters is not surprising - it should be noted that for this particular application over-fitting is not considered to be significantly detrimental to the intended end use.

Finally, the practical approach to achieving multi-class segmentation with sufficient accuracy would seem to be training the network on a large ‘imperfect’ dataset, and then fine-tuning on a small, expertly segmented, dataset.

5 Conclusions

The work presented herein demonstrates that training with a larger, imperfectly segmented dataset outperforms a very small, expertly segmented dataset. Intuitively the small dataset doesn’t provide sufficient examples for the model to learn a general representation of the subject; consequently the model suffers both low accuracy and over-fitting. Furthermore, the larger dataset provides sufficient number of accurate segmentations to overcome the noise. Therefore, for specialist subject matters it is preferable to build a large dataset at the expense of introducing noise to the segmentations. Finally, it is suggested that, in the context of corrosion detection, a viable strategy would be to first train on the large dataset, before fine-tuning on the expert labelled dataset to both improve accuracy and increase discrimination of discrete classes.
6 Acknowledgement

The authors wish to acknowledge the support provided by Woodside Petroleum in providing images of corrosion.

7 Author Contributions

WN prepared the datasets, performed training and analysis of the models, and prepared the manuscript (50%), TD provided technical guidance for the coding and reviewed the manuscript (20%) NB provided technical support for expert labelling of images and reviewed the manuscript (30%).

8 Conflict of Interest

To the authors' knowledge there are no conflicts of interest to declare.
References

Babicky M A 2004 *Psychosomatic Medicine* **66**(3), 411–421.
URL: [http://www.psychosomaticmedicine.org/cgi/doi/10.1097/01.psy.0000127692.23278.a9](http://www.psychosomaticmedicine.org/cgi/doi/10.1097/01.psy.0000127692.23278.a9)

Bearman A, Russakovsky O, Ferrari V & Fei-Fei L 2016 *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9911** LNCS, 549–565.

Cho J, Lee K, Shin E, Choy G & Do S 2015.
URL: [http://arxiv.org/abs/1511.06348](http://arxiv.org/abs/1511.06348)

Dharmasiri T, Spek A & Drummond T 2017 in ‘eprint arXiv:1706.07593’.
URL: [http://arxiv.org/abs/1706.07593](http://arxiv.org/abs/1706.07593)

Everingham M, Van Gool L, Williams C K, Winn J & Zisserman A 2010 *International Journal of Computer Vision* **88**(2), 303–338.

Fritsch J, Kuhn T & Geiger A 2013 *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)* pp. 1693–1700.
URL: [http://ieeexplore.ieee.org/document/6728473/](http://ieeexplore.ieee.org/document/6728473/)

Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li & Li Fei-Fei 2009 *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 248–255.
URL: [http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5206848](http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5206848)

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P & Zitnick C L 2014 *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8693** LNCS(PART 5), 740–755.

Long J, Shelhamer E & Darrell T 2015 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 3431–3440.

Lu L, Shin H c, Roth H R, Gao M, Lu L, Member S, Xu Z, Nogues I, Yao J, Mollura D & Summers R M 2016 *IEEE Transactions on Medical Imaging* **35**(5), 1285–1298.

Menze B H, Jakab A, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber M A, Arbel T, Avants B B, Ayache N, Buendia P, Collins D L, Cordier N, Corso J J, Criminisi A, Das T, Delingette H, Demiralp C, Durst C R, Djoat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin K M, Jena R, John N M, Konukoglu E, Lashkari D, Mariz J A, Meier R, Pereira S, Precup D, Price S J, Reza S M S, Ryan M, Sarikaya D, Schwartz L, Shin H C, Shotton J, Silva C A, Sousa N, Subbauna N K, Szekely G, Taylor T J, Thomas O M, Tutson N J, Unal G, Vasseur F, Wintermark M, Ye D H, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M & Koen Van Leemput 2016 *IEEE Trans Med Imaging* **34**(10), 1993–2024.
URL: [http://www.nature.com/articles/sdata2017117](http://www.nature.com/articles/sdata2017117)

Milan A, Pham T, Vijay K, Morrison D, Tow A W, Lin L, Erskine J, Grinover R, Gurman A, Hunn T, Kelly-Boxall N, Lee D, McTaggart M, Rallos G, Razjigaev A, Rowntree T, Shen T, Smith R, Wade-McCue S, Zhuang Z, Lehnert C, Lin G, Reid I, Corke P & Leitner J 2017.
URL: [http://arxiv.org/abs/1709.07665](http://arxiv.org/abs/1709.07665)

Reale C, Nasrabadi N M & Chellappa R 2016 pp. 1192–1196.

Rolnick D, Veit A, Belongie S & Shavit N 2017 pp. 1–10.

Shwartz-Ziv R & Tishby N 2017 pp. 1–19.
URL: [http://arxiv.org/abs/1703.00810](http://arxiv.org/abs/1703.00810)

Simonyan K & Zisserman A 2014 in ‘ICLR 2015’ pp. 1–14.
URL: [http://arxiv.org/abs/1409.1556](http://arxiv.org/abs/1409.1556)

Taigman Y, Yang M & Ranzato M 2014 *CVPR IEEE Conference* pp. 1701–1708.
Teichmann M, Weber M, Zoellner M, Cipolla R & Urtasun R 2016.  
**URL:** http://arxiv.org/abs/1612.07695

Wei Y, Liang X, Chen Y, Shen X, Cheng M M, Feng J, Zhao Y & Yan S 2015 *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* (Xx), 1–8.  
**URL:** http://arxiv.org/abs/1509.03150