Corrosion detection on aircraft fuselage with multi-teacher knowledge distillation

K. Zuchniak[1,3,*], W. Dzwinel[1], E. Majerz[1], A. Pasternak[1], and K. Dragan[2]

[1] AGH University of Science and Technology., Department of Computer Science, Cracow, Poland
[2] Air Force Institute of Technology, Department of Airworthiness, Warsaw, Poland
[3] Neuralbit Technologies sp. z o. o., Cracow, Poland

Abstract. The procedures of non-destructive inspection (NDI) are employed by the aerospace industry to reduce operational costs and the risk of catastrophe. The success of deep learning (DL) in numerous engineering applications encouraged us to check the usefulness of autonomous DL models also in this field. Particularly, in the inspection of the fuselage surface and search for corrosion defects. Herein, we present the tests of employing convolutional neural network (CNN) architectures in detecting small spots of corrosion on the fuselage surface and rivets. We use a unique and difficult dataset consisting of 1.3×10^4 images (640×480) of various fuselage parts from several aircraft types, brands, and service life. The images come from the non-invasive DAIS (D-Sight Aircraft Inspection System) inspection system, which can be treated as an analog image enhancement device. We demonstrate that our novel DL ensembling scheme, i.e., multi-teacher/single-student knowledge distillation architecture, allows for 100% detection of the images representing the "moderate corrosion" class on the test set. Simultaneously, we show that the proposed ensemble classifier, when used for the whole dataset with images representing various stages of corrosion, yields significant improvement in the classification accuracy in comparison to the baseline single ResNet50 neural network. Our work is the contribution to a relatively new discussion of deep learning applications in the fast inspection of the full surface of an aircraft fuselage but not only its fragments.

Keywords: aircraft maintenance, deep learning, ensemble learning, knowledge distillation, fuselage corrosion detection, DAIS system

1 Introduction

Corrosion, fatigue, and corrosion-fatigue cracking are the most common types of structural problems experienced in the aerospace industry. To ensure flight safety of aircraft structures, it is necessary to have regular maintenance by using visual methods of non-destructive inspection (NDI) [16]. Traditionally, visual

*Corresponding author: Konrad Zuchniak, zuchniak@agh.edu.pl

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

inspections are conducted by human operators that scan the aircraft fuselage looking for corrosion, cracks, and incidental damage. However, this is a costly and time-consuming procedure apt to be subjected to human mistakes caused by mental fatigue and boredom.

In the last decade, various image processing algorithms have been applied in the field of aircraft inspection [18]. However, these algorithms work well only in controlled environments. They often fail in more complex real-world scenarios due to noisy and complex backgrounds. Therefore, used together with the classical machine learning models, fine-tuned image processing techniques are strongly biased by the type of datasets considered.

The success of deep learning in many domains of science and engineering, particularly, the efficacy of various convolutional neural network (CNN) architectures in producing amazingly accurate data models for images (in terms of classification, object recognition, semantic segmentation and others) encouraged us to test this technology as an autonomous support for the inspection system of wide-area surface of the aircraft fuselage. The data for our research come from the imaging acquisition system DAIS (D-Sight Aircraft Inspection System) [11] widely used by the Polish air force and collected by the Air Force Institute of Technology (AFIT). DAIS images are able to enhance the hidden corrosion spots invisible to the naked eye in similar lighting conditions.

To decrease the costs simultaneously increasing the reliability of this time-consuming procedure, herein we propose to support it by the autonomous system based on advanced neural network architectures. From application point of view, the main target of this research is to improve and partially automate aircraft fuselage inspections. Additionally, the research aspect of this work, not directly related to the domain of aircraft inspection, is the use of knowledge distillation as an ensemble learning aggregation mechanism. We can summarize our contributions as follows:

1. We have tested several CNN architectures on DAIS images and estimate their various degree of usefulness in recognition of corroded fuselage rivets.
2. To solve the problems with high data inhomogeneity - data coming from many types of airships of various ages, very typical ones in many inspection/fault detection systems - we propose using the ensemble learning concept to increase generality and to deal with overfitting. Moreover, we modified the knowledge distillation framework [10], to allow its aggregation from the whole $multi - teacher$ ensemble into only one *student* model.
3. We have developed a novel method for mimicking ensemble output by the knowledge distillation employing a multi-teacher/single-student network. This type of knowledge distillation allows training a single CNN of a similar accuracy as the CNN ensemble but requiring more modest resources (i.e. storage size and shorter response time). The experiments show on the test data that it yields superior accuracy among other tested CNNs architectures.

Summing up, we demonstrate that proposed CNN architectures have sufficient classification power to be considered as a valuable support in the wide-area inspection of the aircraft fuselage.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

In the following section, we shortly present the main idea of the DAIS image acquisition system and the dataset that is the subject of our study. Then we describe the methodology proposed, i.e., (1) the ensemble learning scheme and (2) a new knowledge distillation variant based on the multi-teacher/single-student approach. Next, we present a detailed description of the experiments and discuss their results. Our approach and its results can be confronted with the state-of-the-art DL applications in the aircraft inspection in the Related Work section. Finally, we summarize the conclusions and suggest future research goals.

2 Methodology

To deal with overfitting problem, resulting from the relatively small number of examples and the high complexity of our data set, we use ensemble learning. We trained several models whose aggregate predictions were more accurate compared to a single model. On the other hand, the use of ensemble increases computational complexity of our solution, which is also treated as a big disadvantage. To solve this problem, we use the knowledge distillation, transferring the objective knowledge of the entire ensemble to the weights space of only one model. Ensembling models of various types (e.g., formal mathematical models and data models), usually lead to better results, i.e., better approximations, predictions or classification accuracy [20]. However, this is not for free but at the expense of the increase of model storage&time complexity. Therefore, the high demand for computational resources required by big data models (such as ensembled DNNs) is still a challenging problem.

The proposed methodology combines ensembling and knowledge distillation approaches into a new multi-teacher/single-student approach. In the following subsections, we present shortly its principles on the background of ensembling and knowledge distillation techniques.

2.1 Ensemble learning

The main purpose of ensembling the models is to increase the accuracy of predictions [20]. Especially, in the cases of very fine image details and high uncertainty caused by inhomogeneity of data. This is just the case to be encountered in the detection of corrosion on small fuselage rivets from very inhomogeneous data coming from many types of airships of various ages. Thus, the data set consists of many "fuzzy" small subsets having a specific structure (fine graining) while we are looking for common attributes of corrosion (coarse-graining) neglecting fine structural details of data. Therefore, the ensembling of neural nets, each trained on a different homogeneous part of the baseline data set, is an encouraging idea to be applied in the diagnostic of aeronautical structures. The benefits of this DNNs architecture in the context of the aircraft fuselage inspection is presented in the seminal publication [18].

On the other hand, ensemble learning can increase the computational complexity of the predictive model. In the classic implementation of ensembling of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

neural networks (e.g. bagging), each single sub-model is generated in an independent training process. Consequently, N independent models increases the computational complexity of the classifier N times both in the training and inference phase. In general, however, it is not exactly the truth. The sub-nets can be pre-trained in considerably shorter time than a single baseline model [3,6,22] or the architecture of the sub-models can be much simpler than the baseline model.

2.2 Knowledge distillation

The main idea behind the knowledge distillation is that the simpler *student* model mimics the complex *teacher* model resulting with its better interpretation or obtaining a simpler and competitive black box with similar or even superior performance. In this way, the knowledge inscribed in the *teacher* model weights is compressed and transferred into the parameter space of the $student's$ model. This technique was popularized by Hinton et al. in [10]. In the standard training process of a classifier, the loss function is closely related to the data labels. In the case of knowledge distillation, the loss function has a second component related to the distance between *teacher* and *student* output logits. More details and variants of knowledge distillation are presented in the survey paper by Gou et al. [7]. Versatility resulting from the concept of knowledge distillation comes primarily from the lack of requirements for types of the *teacher* and *student* models. This technique is most often used to compress machine learning models based on neural networks whose architectures are very similar, differing only in the number of layers and neurons and weights in each layer. On the other hand, there are no formal requirements as to the type of ML model used. It is possible to distill knowledge between machine learning models of completely different structure, type and principle of operation. It is sufficient to ensure that both models have the same output and input structure.

2.3 Multi-teacher/single-student network

In our approach, we assumed the lack of formal restrictions of knowledge distillation (comparing hidden layers activations requires consistency between *teacher* and *student* architectures). We treat the entire ensemble (composed of sub-nets trained on unique, randomly generated subsets of training dataset) as the *teacher* model. As result, we can use knowledge distillation as an ensemble *decision fusion* scheme. The *student* model learns to mimic predictions of the whole ensemble of the *teacher* sub-models.

There have been several studies that utilize knowledge distillation as ensemble aggregation [2, 25]. However, there are a few important differences between those and our approaches. The major modification assumes to transfer the decision about the aggregation of individual sub-models from the stage before knowledge distillation to the *student* model itself. The main task of the *student* model is not to imitate some aggregation of $teachers'$ outputs, e.g., averaging them, but all individual *teachers* "cooperate" during training synchronously developing a

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

more sophisticated common response. We also decided not to include in the loss function the factor representing the similarity between *teacher* and *student* in internal NN layers. Our loss function forces only mimicking the *teacher* output predictions by the *student* model. This approach gives more flexibility in terms of knowledge distillation between models of a different type (similar architecture is not required), which we plan to use in the future. The temperature parameter T in $softmax$ probabilities

$$P_i = \frac{e^{\frac{y_i}{T}}}{\sum_{k=1}^{n} e^{\frac{y_k}{T}}} \tag{1}$$

is set to 1 what means the loss function utilizes unchanged $softmax$ output.

We analyzed three variants of the multi-teacher/single-student architecture, with different sub-models prediction-aggregation schemes.

1. **Prediction averaging** - Currently used approach [24] consists in averaging the ensemble predictions before the *teacher* output is included in the *student* loss function. The *student* model learns to mimic the average response of the $multi - techer$ ensemble (Fig. 1 upper).
2. **Mimic of prediction geometric center** - In the training process, the output of the *student* model is compared with the predictions of all N $teachers$ individually. The *student* model learns to mimic predictions of several $teachers$ simultaneously. However, since bringing prediction too closely to a single *teacher* output increases part of the loss function responsible for mimicking other $teachers$, *student* model output settles in the geometric center of all the $teachers'$ predictions (Fig. 1 center).
3. **Independent mimicking of all the *N teachers*** - In contrast to the model presented above the *student* model does not produce a single output, but N outputs - where N is equal to the number of $teachers$, each is characterized by an independent set of trainable weights, see Fig. 1 lower. The last layer or the last few layers may be separated. Each of these independent outputs in the training process is compared with its assigned *teacher* output. It should be noted that the convolution part responsible for the feature extraction is common. However, the weights of the last layer (or last few layers) responsible for classifications are specific to each *teacher*. This way, the model does not learn to mimic the aggregation of all *teacher*'s outputs but actually generates N independent predictions linked to each *teacher*.

As shown in [24] there are many loss function definitions, different distance matrices, distillation strategies *et cetera*. We decided to use hard *ground true* labels and hard ensemble outputs, instead of light labels (in which the labels do not have the entire probability assigned to one class, but it is partially "fuzzified" to other classes). We also used the Kullback-Leibler divergence (KLD) to determine the distance between *teacher* and the *student* models. Below we present respective equations determining loss function for $teacher - student$ models variants described above.

$$Loss_{avg} = \alpha \sum_{i=1}^{D} \bar{y_i} \cdot log(\frac{\bar{y_i}}{\tilde{y_i}}) \; - \; (1-\alpha) \cdot \sum_{i=1}^{D} y_i log(\tilde{y_i}) \tag{2}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

$$Loss_{geo} = \alpha \frac{1}{N} \sum_{i=1}^{D} \sum_{j=1}^{N} y_{ij} \cdot log(\frac{y_{ij}}{\bar{y_i}}) \; - \; (1-\alpha) \cdot \sum_{i=1}^{D} y_i log(\tilde{y_i}) \tag{3}$$

$$Loss_{ind} = \frac{1}{N} \sum_{j=1}^{N} \cdot (\alpha \sum_{i=1}^{D} y_{ij} \cdot log(\frac{y_{ij}}{\tilde{y_{ij}}}) - (1-\alpha) \cdot \sum_{i=1}^{D} y_i log(\tilde{y_{ij}})) \tag{4}$$

,

where D is the *student* output size (number of classes), N is number of *teachers*, $\tilde{y_i}$ is the i-th scalar value in the *student* model output, y_i is the corresponding target value, $\bar{y_i}$ is the corresponding average of *teachers* model output, y_{ij} is the corresponding j-th *teacher* output, and $\tilde{y_{ij}}$ is the i-th scalar value in j-th output of *student* model output (independent mimicking variant). The α weight setting proportion between the expression associated with knowledge distillation (first sum in equations) and the standard loss function connected with data *ground truth* (second sum), is the process controlling parameter, increasing this parameter, increases student imitation loss in the total loss function. Figure 1 demonstrates the block diagrams of all multi-teacher/single-student networks described above in the order presented in the text.

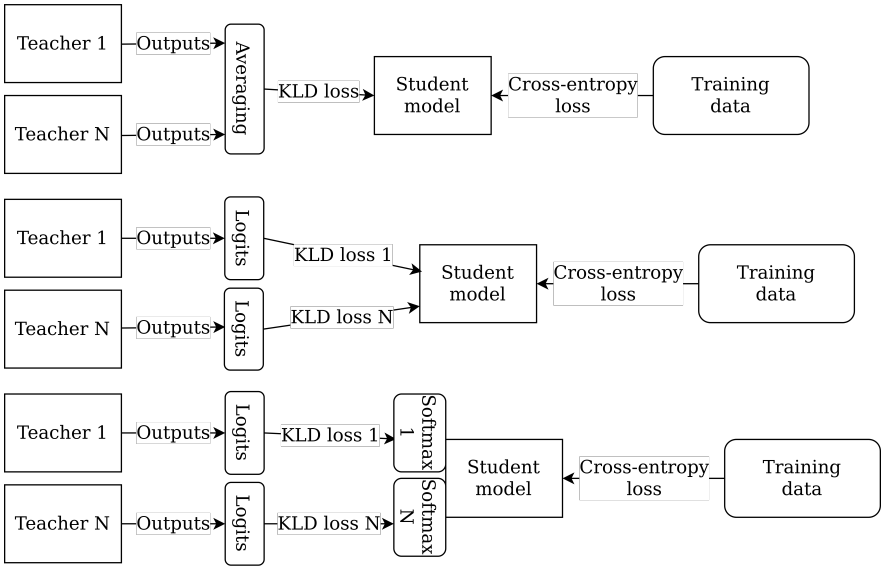

Fig. 1. The schemes of the multi-teacher/single-student models employed in this paper: *prediction averaging* model (upper), the model *mimic of prediction geometric center* (middle), *independent mimicking of all the N teachers* (bottom).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

3 DAIS data

3.1 DAIS System

D-Sight [9, 13] is an optical double-pass retroreflection surface inspection technique created by Diffracto Ltd from Canada. It is a patented method of visualizing very small surface distortions outside the plane, such as dents and corrosion. The D-Sight optical system consists of a retroreflective screen, camera, a light source, and a tested fuselage fragment (Fig. 2). The light from a standard divergent source is reflected off the sample. The surface of the sample must be reflective. The reflected light is then shone onto a reflective screen, which consists of many semi-silvered glass spheres (typical diameter 60 µm). This screen tries to redirect all incident light rays at the same angle to the starting point of reflection on the sample surface. However, the screen is not perfectly reflective and actually returns a divergent cone of light rather than a single beam at the same angle. It is this imperfection of the reflective screen that creates the D-Sight effect. The light is reflected again by the sample and collected by a camera slightly away from the light source.

DAIS system [5] uses this imaging technology for damage detection which are not visible to the naked eye. Figure 2 presents the overview drawing containing the principle of operation for the DAIS imaging system and a photo showing the process of fuselage image acquisition.

The detection system of corrosion on the aircraft fuselage consists of many modern non-invasive visual inspection techniques presented in [17] including also a highly modernized, comparing to its original version, imaging tool based on D-Sight methodology.

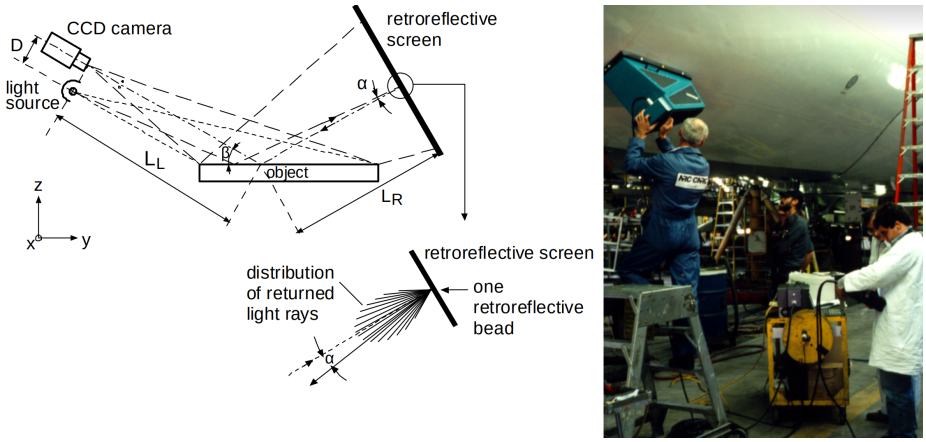

Fig. 2. Left: a scheme of DAIS imaging system operation, from [9] Right: the image demonstrating the process of captioning aircraft images.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

3.2 Fuselage corrosion dataset

Thanks to the Air Force Institute of Technology (Warsaw, Poland), we got access to data representing the images acquired by using D-Sight technology. We received about 1.3×10^4 labeled images (640×480 pixels). The labels include the testing year, the anonymous *id* of the aircraft, and the label representing the extent of corrosion damage.

Figure 3 shows the data details, i.e., the frequency distribution of samples according to the year of technical examination and an aircraft *id*. Sample images from the DAIS system are also shown. We aim to classify the images according to the strength of the identified damage. Due to the imbalanced data set, we decided to consider this problem as a binary classification: "no damage" and "damage detected". The original images come in 640×480 resolution, however, following the guidelines of the authors of the models used, we have reduced the resolution to 320×240 - for training and inference speed up. The tests, carried out while training the models at full resolution, showed a minimal decrease in the classification accuracy [15].

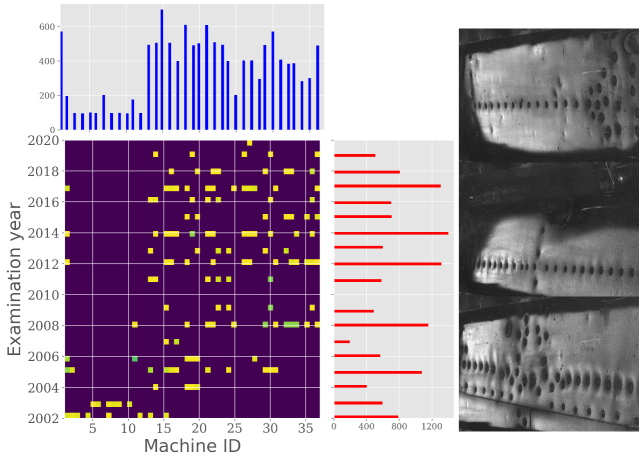

Fig. 3. Left: Distribution of the image examples in DAIS dataset with machine id and inspection year, Right: DAIS samples. Top: no corrosion, Center: light corrosion, Bottom: moderate corrosion. Total number of examples in the data set by class: no corrosion and no damage:6431, light corrosion:6040, moderate corrosion:578, strong corrosion:0, minor damage:26. Histograms presents marginal distribution according to machine ID and examination year dimension.

4 Results and discussion

4.1 Hardware and software setup

The computations were performed on the Prometheus supercomputer (288th on top500 list (June 2020); HP Apollo 8000, Xeon E5-2680v3 12C 2.5GHz, Infini-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

band FDR, HPE Cyfronet Poland). We used just one node (Intel Xeon E5-2680 v3, 2.5 GHz) and two Nvidia V100 GPU accelerators on the cluster dedicated to deep learning. In the computations We used TensorFlow framework [1].

4.2 Experiment description

Using previous analyzes, we have selected ResNet50 [8] as the baseline CNN architecture. We show in the supplementary materials [15] that this architecture produces the best and more stable results comparing to the others.

ResNet50 training setup are as follows: ADAM optimizer [12], learning rate = 0.001, batch size = 128, number of epochs = 150. The dataset was split into training, validation, and test parts, based on the aircraft *id*. Samples from machines with *id* between 1 and 30 were assigned to the training data set (10 534 examples), from *id* = between 31 and 34 were assigned to the validation set (1463 examples) while samples from aircraft with *id* between 35 and 37 were assigned to the test set (1297 examples). Completely different physics of acquiring images from the DAIS system compared to standard photography was reason of resigning from use of transfer learning. Our experiments have shown that use of pre-trained models does not improve the quality of classification on DAIS data, we test models trained on ImageNet100 [4], and we achieve lower classification accuracy.

The ensemble classifier consists of ResNet50 sub-nets (*teachers*) was trained on different training subsets. We generated many ResNet50 sub-nets, each trained on different, randomly generated subset of training data. The examples were generated in such a way that the percentage of common examples for any two selected subsets was the same. Thanks to this approach, we obtain the maximum diversity of generated data subsets. In the next step we applied knowledge distillation to aggregated (trained) ensemble of *teachers* into a single *student* model. We chose a number of sub-models $N = 5$ and *coverage* factor equal to 0.7 (defined as the size of the training subset relative to the entire training set).

4.3 Corrosion detection

We tested and compared three main approaches (based on ResNet50 architecture) described in previous section in order to develop corrosion classifier. Depending on the threshold level, We can modify the trade-off between the number of false negatives and false positives. Figure 4a shows precision and recall metrics depending on the threshold value. We define the threshold as the minimum value of the probability of assigning a sample to the "corrosion detected" class. Images labeled as "corrosion detected" came from several more specific classes representing various degrees of material failure. We conducted the analysis for these specific corrosion classes, comparing how the models dealt with samples labeled as "light corrosion" and "moderate corrosion". The results appeared to be very promising. On the test set our models were able to recognize 100% "moderate corrosion" samples (with the appropriate threshold level). Unfortunately,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

the test set of examples with "moderate corrosion" is limited to only 79 examples. On the other hand, from the application point of view, the detection of stronger examples of corrosion is the most important and can be the positive test for the usefulness and reliability of our detection algorithm. For safety reasons, in the operation of the autonomous corrosion detection system, the detection of stronger corrosion samples is crucial. It should also be remembered that the whole data set was manually labeled by experts and this may be the reason for the existence of some bias (incorrect markings for pairs "no corrosion" - "light corrosion"). Figure 4b demonstrates different recall curves for specific corrosion levels.

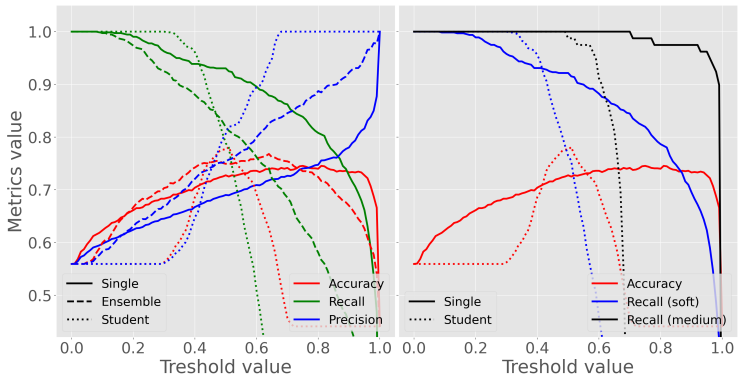

Fig. 4. Left: Precision-recall characteristics for the models considered. The intersection of recall and precision lines is at the highest point for the *student* model. Right: Precision-recall characteristic with separation for "light" and "moderate" corrosion levels. The "moderate corrosion" samples are much better recognized by the models. We achieved 97.5%-100% detection of corrosion on this level.

To determine the appropriate threshold values for a fair comparison of the various methods, we assessed them independently for each model. The maximum classification accuracy achieved on the validation set was the selection criterion for the thresholds. Then we calculated the remaining metrics on the test set. For the thresholds selected in this way, the geometric *student* model achieves detection of 100% "moderate corrosion" class while the ensemble and single models get 97.5%. It gives also superior results on the other metrics. The results are collected in Table 1.

To visualize, which areas of analyzed images influence the decision on corrosion classification we use the Grad-CAM [21] method. The algorithm employs the cumulative gradients calculated in back-propagation which are treated as "weights" to explain network decisions. It can be seen that the greatest activations are generated on the riveting line (see Figure 5). As hidden corrosion occurs on the rivets, so this behavior of the model shows a good level of data understanding.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

Table 1. Accuracy, recall, precision and F1 score matrices obtained by tested classifiers. Complexity* is expressed as a relative value, where 1 means complexity level of a single ResNet50 base model

Used model	Threshold	Accuracy	Recall	Precision	F1 Score	Com*
Single	0.89	73.6%	73.96%	77.75%	75.81%	1
Ensemble	0.62	76.25%	74.81%	81.24%	77.89%	5
Averaging student	0.54	74.14%	69.63%	81.43%	75.07%	1
Geometric student distillation	0.47	**76.63**%	**84.26**%	76.39%	**80.13**%	1
Multi output student	0.51	74.3%	69.78%	**81.6**%	75.23%	1

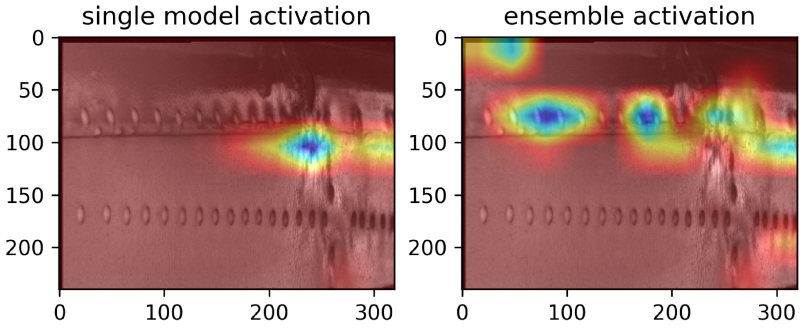

Fig. 5. Grad-CAM activations for a single baseline model (left) and the ensemble (right) model. The activation map for ensemble is much wider. From the explainability point of machine learning models, we can determine that ensemble takes more factors into account when generating predictions.

Additionally, we used t-SNE [14] data embedding method to visualize the localization of samples from the test set in 2D space. The *Single* model and the geometric *student* model were compared. The feature vectors are collected from the output of the global max-pooling layer, which follows the last convolutional layer. The resulting feature vector had 2048 dimensions. Figure 6 shows this feature vector embedding into a 2D space for visualization purposes. It is easy to observe a strong separation between the‘"moderate corrosion" and "no corrosion" classes. The "light corrosion" class lies in the middle area and partly overlaps "no corrosion" class. This result coincides with the classification metrics achieved by the model for individual classes. Data points were normalized to better cover plot canvas. We calculated Silhouette coefficient [19] (*Single*: .0015, *student*: .0359) to quantitatively show that *student* produce better clustering (higher coefficient score means better clusters class separation).

5 Related works

Very few works on aircraft fuselage inspection using modern DNN architectures can be found in the literature. In [23] a deep learning-based framework is proposed for automatic damage detection in aircraft engine borescope inspection.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

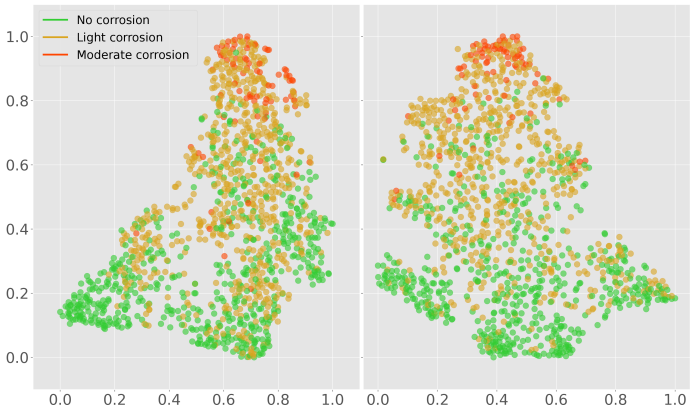

Fig. 6. DAIS samples embedding by using t-SNE. The *Single* model and the geometric *student* model were compared, respectively

It utilizes the state-of-the-art NN model called Fully Convolutional Networks (FCN) to identify and locate damages from borescope images. This framework can successfully identify two major types of damages, namely cracks and burn, and extract ROI regions on these images with high accuracy. In [16] the authors present the system for crack detection on the aircraft fuselage based on high-resolution drone images.

The similar, in spirit, work to that presented here is described in [18]. The authors had a modest dataset consisting of images from a borescope inspection of aircraft propeller blade bores. Due to the limited size of the data, they used the transfer learning by pre-training a convolutional neural network on the large ImageNet, assuming that the low-order features will be the same for both datasets. It is shown that the ensemble method improves inspection accuracy over conventional single CNN. However, the borescope is designed to assist visual inspection of narrow, difficult-to-reach cavities but not to cover big areas of aircraft fuselages. Thus, the data used in this system are completely different from those considered in this paper. However, the success of the application of CNN's ensemble was the inspiration of our paper.

6 Conclusions and future work

In the research presented, we used the state-of-the-art machine learning models to automate the task of corrosion detection on the aircraft fuselage. The images we analyzed were taken from the DAIS imaging system, but we believe that this methodology can be also applied with other visual inspection systems. One of the problems we encountered in this study was the severely limited collection of training data. Moreover, the domain of this data differed significantly from popular image repositories, which meant that transfer learning cannot produce satisfactory results. We have developed a method of aggregation and compres-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

sion of knowledge derived from several machine learning models. The proposed variant of linking the *student's* model loss function with the geometric center of *teachers* ensemble outputs produced the best results in the context of corrosion classification efficacy, giving F1 statistics equal to 80%, i.e., 4,4% more than by employing the single baseline ResNet50 model. Moreover, the images of the most corroded fuselage parts were recognized with 100% accuracy. Supplementary material and the codes we used in our experiments are published here: (https://github.com/ZuchniakK/DAISCorrosionDetection).

In the nearest future, we intend to expand our dataset, which will allow us to generate more accurate models and perform better and more certain validation. We also plan further work on the NN model compression and quantization to enable its implementation directly in DAIS hardware. In the future we intend to test our proposed multi-teacher ensembling framework on the other difficult data sets such as medical images.

The *Geometric student distillation* method we proposed generated the best performing models, but the differences are small. We believe that student generation methods can still be improved and will be the subject of our further research.

Acknowledgements

The research was supported by the funds assigned to AGH University of Science and Technology by the Polish Ministry of Science and Higher Education and in part by PL-Grid Infrastructure. We thank dr Jerzy Komorowski (former General Manager for National Research Council Canada) and professor dr Stan Matwin (Director of Institute for Big Data Analytics, Dalhausie University) for their contribution to this research.

References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), http://tensorflow.org/, software available from tensorflow.org
2. Asif, U., Tang, J., Harrer, S.: Ensemble knowledge distillation for learning improved and efficient networks. arXiv preprint arXiv:1909.08097 (2019)
3. Bukowski, L., Dzwinel, W.: Supernet–an efficient method of neural networks ensembling. arXiv preprint arXiv:2003.13021 (2020)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Forsyth, D., Komorowski, J., Gould, R., Marincak, A.: Automation of enhanced visual ndt techniques. In: Proceedings 1st Pan-American Conference for NDT, Toronto, Canada. pp. 107–117 (1998)
6. Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D.P., Wilson, A.G.: Loss surfaces, mode connectivity, and fast ensembling of dnns. In: Advances in Neural Information Processing Systems. pp. 8789–8798 (2018)
7. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. arXiv preprint arXiv:2006.05525 (2020)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Heida, J., Bruinsma, A.: D-sight technique for rapid impact damage detection on composite aircraft structures (1997)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Jerzy P. Komorowski, Ronald W. Gould, D.L.S.: Application of diffracto sight to the nondestructive inspection of aircraft structures. Review of Progress in Quantitative Nondestructive Evaluation **12**, 449 (2012)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Komorowski, J.P., Gould, R.W., Simpson, D.L.: Synergy between advanced composites and new ndi methods. Advanced Performance Materials **5**(1-2), 137–151 (1998)
14. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
15. Majerz, E., Pasternak, A.: System for analyzing damage to the surface of aircraft stryctures using convolutional neural networks. Bachelor's thesis, AGH University of Science and Technology (2021), https://github.com/majerzemilia/engineering-thesis
16. Malekzadeh, T., Abdollahzadeh, M., Nejati, H., Cheung, N.M.: Aircraft fuselage defect detection using deep neural networks. arXiv preprint arXiv:1712.09213 (2017)
17. Niepokólczycki, A., Leski, A., Dragan, K.: Review of aeronautical fatigue investigations in poland (2013-2014). Fatigue of Aircraft Structures **2016**(8), 5–48 (2016)
18. Ren, I.: An Ensemble Machine Vision System for Automated Detection of Surface Defects in Aircraft Propeller Blades. Ph.D. thesis, Georgia Institute of Technology (2020)
19. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
20. Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1249 (2018)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
22. Sendera, M., Duane, G.S., Dzwinel, W.: Supermodeling: the next level of abstraction in the use of data assimilation. In: International Conference on Computational Science. pp. 133–147. Springer (2020)
23. Shen, Z., Wan, X., Ye, F., Guan, X., Liu, S.: Deep learning based framework for automatic damage detection in aircraft engine borescope inspection. In: 2019 International Conference on Computing, Networking and Communications (ICNC). pp. 1005–1010. IEEE (2019)
24. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. arXiv preprint arXiv:2004.05937 (2020)
25. Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. In: Advances in neural information processing systems. pp. 7517–7527 (2018)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_25