A Comparative Study of Open Source Computer Vision Models for Application on Small Data: The Case of CFRP Tape Laying

Anonymous Authors[1]

Abstract

In the realm of industrial manufacturing, Artificial Intelligence (AI) is playing a increasing role, from automating existing processes to aiding in the development of new materials and techniques. However, a significant challenge arises in smaller, experimental processes characterized by limited training data availability, questioning the possibility to train AI models in such small data contexts. In this work we explore the potential of Transfer Learning to address this challenge, specifically investigating the minimum amount of data required to develop a functional AI model. For this purpose, we consider the use case of quality control of Carbon Fiber Reinforced Polymer (CFRP) tape laying in aerospace manufacturing using optical sensors. We investigate the behavior of different open-source computer vision models with a continuous reduction of the training data. Our results show that the amount of data required to successfully train an AI model can be drastically reduced, and the use of smaller models does not necessarily lead to a loss of performance.

1. Introduction

1.1. From Big Data to Small Data: AI's New Frontier in Manufacturing

Artificial intelligence technology has become increasingly important in industrial manufacturing in recent years. They enable the automation of processes such as quality assurance, predictive maintenance and optimization of production processes. Since powerful AI models require a corresponding big amount of data, the focus at the beginning of this development was primarily on relatively few use cases that provided big data. Highly scaled processes,

such as industrial series production, were particularly suitable for this purpose. Now that this area of application has reached a certain level of maturity, research has recently focused on the far more common applications where only a small amount of data is available. These applications can be found, for example, in the development of new materials and production processes, i.e. in the corresponding pre-series production. While the big data problems were about reducing the dataset by filtering to increase model quality, small data problems raise the question of whether there is even enough data to create an efficient AI model?

1.2. Optimizing CFRP Tape Laying Processes

Let us illustrate this using the example of CFRP tape laying as shown in figure 1. CFRP tape laying is a versatile and efficient manufacturing method prominently utilized within the aerospace and automotive sectors. Its primary purpose is the construction of intricate and robust composite components, where lightweight, high-strength materials are essential. Its great versatility, allows manufacturers to produce parts with complex shapes and tailored mechanical properties. This new technology is on the edge of industrial maturity. This means that it operates in the area of pre-series production. Only small quantities will be produced under variable production processes, as these are subject to continuous optimization. This means that only little data is available for training AI models, such as for quality control. Nevertheless, automated and non-destructive testing is of great importance for these small batches in order to carry out further practical tests. Optical images are currently being taken continuously to detect component defects. The resulting images are further analyzed using complex algorithms. However, these have to be constantly adapted to the ever-changing process conditions during development, which represents a considerable effort. The challenge of this computer vision problem is to provide an automated quality control system that can adapt quickly to changing process conditions. A corresponding AI model would have to be able to adapt to the new process on the basis of a small number of examples classified by humans. Again, the question arises how many examples are necessary to create an AI model?

---

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.
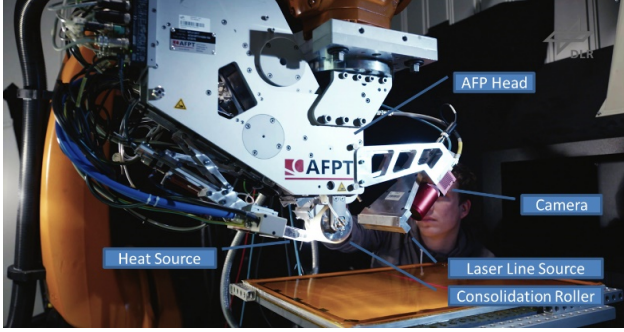
Figure 1: Automated Head for CFRP tape laying with optical sensor for quality control (Schuster et al., 2020).

1.3. Related Work

Transfer Learning in Computer Vision Transfer learning is a common practice in the field of deep learning for computer vision. This technique involves creating high-quality models from large datasets and then fine-tuning them on smaller datasets (Kolesnikov et al., 2019; Djolonga et al., 2020; Dosovitskiy et al., 2021). These models perform significantly better than those trained solely on the smaller datasets. This approach leverages the knowledge gained from large datasets to enhance learning efficiency and accuracy on limited data. It has been widely adopted for computer vision tasks, including image classification, object detection, and, more recently, in complex scenarios such as autonomous driving and medical image analysis.

Deep Learning for CFRP The deep learning method has already been investigated for automatic quality assurance in our CFRP use case (Meister et al., 2021). This exploration addressed the issue of limited data by utilizing Deep Convolutional Generative Adversarial Networks (DCGAN) and geometrical transformation techniques through synthetic image data augmentation. The resulting model is intended to serve as a benchmark for the quality of our fine-tuned models. However, we explicitly choose not to follow this path so that our findings can be applied to other use cases. Instead, we aim to determine under what circumstances and how much data are required to produce a comparable, if not superior, model using transfer learning.

1.4. Our Approach

Objectives and Contributions We want to apply Transfer Learning for the application of optical quality control for the prediction of component defects in CFRP tape laying and show that transfer learning under the given small data leads to powerful models that outperform related solutions in this area. We restrict ourselves here to the use of open source pre-trained models. This should ensure that the present results can be adapted to similar applications and that the

required computing capacities can also be applied to smaller development projects. In our comparative study, we address two key questions in particular. On the one hand, we want to compare different such pre-trained models as a basis for transfer learning. Thus, we want to find out what influence the model architecture and model size has in our use case. On the other hand, we address the question how a reduction of the data available for training affects the performance of the model. In doing so, we want to determine a lower limit for the amount of data required to create an AI model with the desired accuracy.

Outline This is done as follows: In section 2 we describe the use case from a technical point of view and in particular how the data for optical quality testing is generated and what properties it has. Next, in section 3, we explain the methods used in transfer learning, i.e. from the choice of pre-trained models to the preprocessing of the data and the hyperparameter search used. The results obtained are presented and discussed in section 4. Finally, we summarize the main aspects of our work in section 5 and give an outlook on the resulting possibilities.

2. CFRP Tape Laying

Thermoplastic tape lay-up is a process where continuous strips of thermoplastic material, combined with reinforcing fibers such as carbon, are carefully placed onto a surface. This method is used to build layered composite structures. The tapes are often heated to their melting point to facilitate bonding between layers and the substrate.

2.1. Horizontal Defects: Gaps and Overlaps

During this process, a wide variety of possible defects can occur. In this work we concentrate us on two defects in horizontal direction which are gaps and overlaps, as shown in figure 2. Gaps occur when there is insufficient material coverage between tape strips, while overlaps happen when tape strips excessively cover one another. Both defects may lead to potential structural weaknesses.

2.2. Measurement Techniques in Automated Fibre Placement

Tapelaying and measurement are performed by a specialized robotic end-effector known as the automated fibre placement (AFP) head, as shown in figure 1. This head feeds multiple carbon fibre tapes to a heat source and then to a consolidation roller, where the material is firmly adhered to a jig of the desired final shape. Directly after the initial cool down, a tape placement sensor (TPS) acquires the height profile of the placed tapes by laser triangulation. Every profile, scanned at a distance of 400 microns, is added to a 16-bit gray scale TIFF image (figure 2). The real world
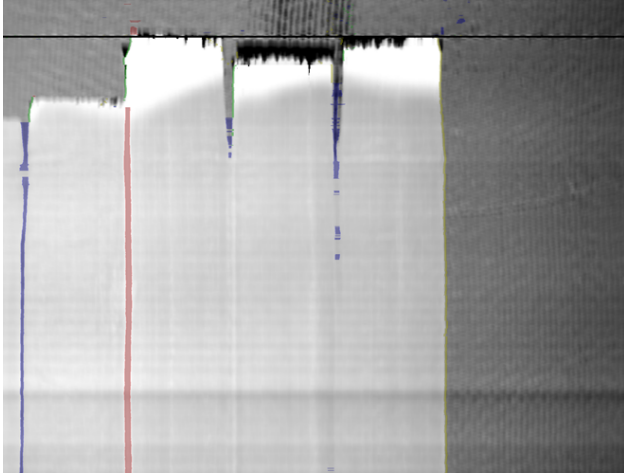
Figure 2: An exemplary high-resolution 16-bit TIFF image with annotation showing the measured height profiles at high contrast. The middle of the pictures shows the actual track consisting of three tapes arranged next to each other, displayed in light gray. To the left is the previous track, to the right the flat dark gray background. Each tape is 12,54 mm wide and may reach several meters in length, depending on the layup. The upper section of the image shows the beginning of the laid tape. On the left side in the previous track between the middle and right tape there is a gap (highlighted in blue), and at the border between the previous and the actual track there is an overlap (highlighted in red). Each of the 65,536 grayscale values of the original TIFF corresponds to a height differential of 1.77 micrometers, i.e., to represent the height difference of a tape, 79 grayscale values are required (Schuster et al., 2020).

height profile may be obtained by scaling the TIFF-image with respect to the optical resolution in x-direction, the scan distance in y-direction and the sensor height calibration in z-direction. The previously described defects 'gap' and 'overlap' are then marked as mentioned in section 1, the algorithm can be manually adjusted until the desired accuracy is reached.

2.3. Dataset Overview: Composition and Imbalance

In order to cover a sufficient defect size for the AI approach. Images of 152 pixels by 100 pixels were generated to cover each category. The pixels correspond to a physical image section of 6,2 mm by 40 mm. Due to a 0,4 mm resolution along the layup direction and 0,041 mm along the image width. Consequently, a dataset with 73,749 examples is formed, comprising 8,707 with a gap, 3,248 with an overlap, and 61,794 without defects, designated as nominal. The nominal category dominates with approximately 84%, followed by gap with 12% and overlap with 4%. Thus, we are dealing here with an unbalanced dataset in which the nominal examples clearly dominate the two error categories. This corresponds to the real use case in industrial production, where a defect should be the exception and not the rule.

3. Computational Methods

As outlined in Section 2, the height profiles of tape layers, as captured in the process, are interpretable as image data. Our goal is to develop an AI model with supervised machine learning that has the task of classifying these images into the error categories "gap", "overlap" and "nominal", as previously described in section 2. This is a classic challenge for image classification in the field of computer vision. A variety of pre-trained models are available for this purpose.

3.1. Pre-Trained Models

Deep Learning in Computer Vision remains a dynamic research area, with state-of-the-art models continuously evolving. For a specific use case, it is challenging to predict the most promising architecture (Zhou et al., 2021). In order to reflect the current diversity, we will choose our pre-traines models from the following architectures:

Convolutional Neural Networks (CNNs) The 2010s were a period of groundbreaking advances in computer vision, initially reflected in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) leaderboards for object recognition and image classification (Russakovsky et al., 2014). CNNs initially proved to be a superior approach, like VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016) and AlexNet (Krizhevsky et al., 2012). Inspired by the biological function of the visual cortex (Hubel & Wiesel, 1962), the fundamental architecture was established in the 1980s and has been continuously developed (Fukushima, 1980; LeCun et al., 1989; Lecun et al., 1998). Due to their computationally intensive nature, a breakthrough was achieved only with the adaptation of the algorithms to the powerful GPUs available in the 2010s (Cireşan et al., 2010; 2011).

Vision Transformers (ViT) ViTs represent another promising approach in deep learning. Initially designed for text processing (Vaswani et al., 2017), the core concept of the attention mechanism was successfully adapted for image analysis. These resulting ViTs are adept at recognizing global structures in images and have swiftly outperformed state-of-the-art ResNet-based baselines in challenging datasets (Dosovitskiy et al., 2021).

[1] Reference for Hugging Face Model Hub (Hugging Face, 2024).

[2] Size of PyTorch model binary file.

3

Table 1: Overview of Selected Models with Corresponding Hyperparameters.

Short Name	Checkpoint[1]	Reference	Model Size[2]	Batch Size	Learning Rate
beit-base	microsoft/beit-base-patch16-224	(Bao et al., 2021)	350	32	3.80E-06
beit-large	microsoft/beit-large-patch16-224	(Bao et al., 2021)	1259	64	3.72E-05
deit-tiny	facebook/deit-tiny-patch16-224	(Touvron et al., 2021)	23	16	3.63E-05
deit-base	facebook/deit-base-patch16-224	(Touvron et al., 2021)	346	16	4.14E-05
dinov2-small	facebook/dinov2-small	(Oquab et al., 2023)	88	16	5.61E-06
focalnet-tiny	microsoft/focalnet-tiny	(Yang et al., 2022)	114	16	5.61E-06
focalnet-base	microsoft/focalnet-base	(Yang et al., 2022)	353	32	8.41E-05
resnet-18	microsoft/resnet-18	(He et al., 2016)	46	16	3.63E-05
resnet-50	microsoft/resnet-50	(He et al., 2016)	103	16	9.26E-05
resnet-101	microsoft/resnet-101	(He et al., 2016)	167	16	9.26E-05
vit-base	google/vit-base-patch16-224-in21k	(Wu et al., 2020)	1198	32	1.15E-05
vit-large	google/vit-large-patch16-224-in21k	(Wu et al., 2020)	1249	64	2.78E-05

Focal Modulation Networks (FMNs) Recent developments have introduced a novel approach, replacing the transformer's self-attention mechanism with a focal modulation module that leverages convolutional layers for modeling token interactions in vision (Yang et al., 2022). This FMN architecture ranks among the latest state-of-the-art models, demonstrating superior performance over modern attention-based models like BEIT-3 (Wang et al., 2023) on the COCO dataset (Lin et al., 2014).

Open Source As motivated in section 1, we exclusively utilize pre-trained open-source models. In particular, we employed models from the Hugging Face Model Hub (Hugging Face, 2024). The selected model names, architectures, and checkpoints are detailed in Table 1.

3.2. Data Preprocessing

Training, Evaluation and Test Data Our dataset is randomly partitioned into a training set and a test set. We allocate 70% of the data for training purposes, while reserving 30% to uniformly assess the performance of the resulting model. Furthermore, from the training data, an additional 10% is set aside for evaluation purposes during the training process. This approach ensures that the random split maintains the distribution of categories as described in section 2.

Balancing As described in section 2, the dataset has an unbalanced distribution. To prevent the model from favoring the dominant category, the training set is balanced. We simply select an equal number of examples from each class, i.e. "nominal", "gap" and "overlap", to ensure a more even representation across all categories. Thus, we improve the classifier's ability to make accurate predictions across all categories without overfitting to the most prevalent category.

Image Preprocessing The pre-trained models described in Table 1 were pre-trained on color image datasets, such as ImageNet (Russakovsky et al., 2014). Consequently, they typically expect three two-dimensional 8-bit arrays of a specific size, commonly 224x224 pixels. These three channels correspond to the red, green, and blue color channels, with each 8-bit channel capable of representing values ranging from 0 to 255. However, the images from the dataset described in Section 2 are 16-bit grayscale images of size 150x100 pixels. These images have only one channel with values ranging from 0 to 65,536. To apply the models to our dataset, the images need to be appropriately pre-processed. First, the grayscale TIFF images are converted into two-dimensional 8-bit arrays. During this process, no normalization is performed, which would otherwise distort the physical height profile of the images. Instead, the average value of individual pixels is subtracted from all pixel values, and values below 0 or above 255 are appropriately capped. In a second step, this array is duplicated, and the required three color channels are created. Finally, the images must be resized to match the input size of the models. The models we selected expect images of 224x224 pixels or larger. Again, we avoid scaling to prevent distortion of the physical information in the images. Instead, the images are padded with zeros at the edges to reach the required size. We can particularly forego normalization in this case, as the height difference between the substrate and the top surface of the tape laying, as described in Figure 2, can be represented within the 255-pixel range.

Data Augmentation Despite the limited size of our dataset, particularly in the smallest category examples, we have chosen to forego additional data augmentation methods. This decision is based on two considerations: Firstly, we aim to ensure the transferability of our results to other

application cases where data augmentation may not be an option. Secondly, there are studies indicating that aggressive data augmentation techniques, such as noising, flipping, and rotation, can significantly impact the performance of pre-trained models (Zoph et al., 2020).

3.3. Hyperparameter Search

For each selected model, a hyperparameter search is conducted to optimize the choice of learning rate and batch size during training. To reduce computational effort, two measures are implemented. Firstly, this is conducted only once beforehand for each model. For this purpose, the training dataset introduced is reduced to 512 examples per category, while the evaluation dataset is retained. Secondly, we employ the Ray Tune (Liaw et al., 2018) implementation of the Asynchronous Successive Halving Algorithm (ASHA) (Li et al., 2018) for hyperparameter tuning. The search space for the hyperparameter search is described in Table 2, and the settings for ASHA are in Table 3. The results of the hyperparameter search are presented in Table 1.

Table 2: Parameters for Hyperparameter Search Space.

Parameter	Values
Learning Rates	[1e-6, 1e-4]
Batch Sizes	[16, 32, 64, 128[1]]

Table 3: Parameters for Ray Tune's (Liaw et al., 2018) implementation of Asynchronous Successive Halving Algorithm (ASHA) (Li et al., 2018).

Parameter	Values
Max T	32
Grace Period	4
Reduction Factor	2
Number of Trials	64

3.4. Fine-Tuning and Evaluation

Using the hyperparameters listed in Table 1, we train the selected models on a balanced dataset that includes 200, 400, 600, 800, 1200, 1600 and 2000 instances for each category "gap", "nominal" and "overlap". Following the hyperparameter optimization procedure, the evaluation dataset is kept without modification. After fine-tuning, we evaluate the performance of the models using the previously described test dataset. Since our problem is a multi-class problem, we calculate the mean F_1 score to evaluate the classification efficiency. The outcomes of these calculations are illustrated in Figure 3.

---

[1]The search was conducted on three Nvidia A6000 GPUs with two trials per GPU. For larger models, this batch size was removed due to memory constraints.

4. Results

We aim to evaluate the performance of the fine-tuned models from Table 1 across varying sizes of training datasets, that is, under a defined number of available examples per category 'gap', 'nominal', and 'overlap'.

4.1. Impact of Architecture

To evaluate the impact of the architecture on the performance of the fine-tuned models, in Figure 3 we investigate the F_1 score of each model in relation to the provided examples per category.
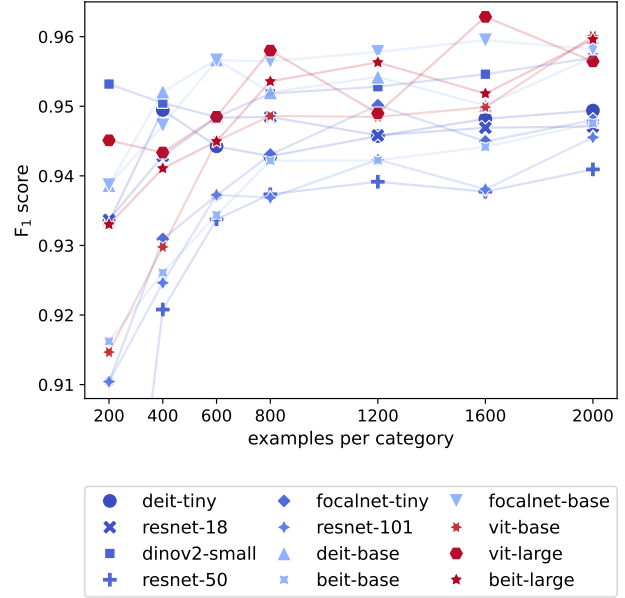

Figure 3: Analysis of sensitivity to training dataset size for fine-tuned models from table 1: Averaged F_1 score versus number of examples per category in the training dataset. The results are colored in relation to the PyTorch model binary file sizes in a continuous color gradient from blue (small) to red (large).

Area of Stable Performance We observe that under the largest dataset with 2000 examples, all models show a good F_1 score. The weakest model, the CNN ResNet-18, scores approximately 0.94 points, only about 0.02 points below the best model, vit-base, which scores around 0.96 points. Upon closer differentiation, a leading group emerges, showing only a 0.005 point difference among themselves. This group exclusively includes ViTs (dinov2-small, vit-base, vit-Large, deit-base, and beit-large) and one FMN (focalnet-base). A chasing group of models with ViTs (deit-tiny, beit-base), FMNs (focalnet-tiny), and CNNs (resnet-18, resnet-50, resnet-101) is clustered about 0.01 point below. With a

continuous reduction in the number of examples per category available for fine-tuning, these two groups form relatively well-separated bands with stable performance down to a threshold of 800 examples.

Area of Performance Decline   When the threshold of 800 examples is crossed, nearly all models experience an accelerating decline in the F_1 score. For the leading group, this decline is less pronounced at approximately 0.02 points, compared to the following group, which sees a decline of approximately 0.03 points. The groups largely remain intact, with only the two ViTs deit-tiny and vit-base and the CNN resnet-18 changing bands. The ViT dinov2-Small proves to be particularly stable, showing no significant loss in comparison to having 200 examples.

4.2. Influence of model size

The previously mentioned transition of resnet-18, the smallest CNN in our selection, from the chasing to the leading group. We aim to examine the different behaviors of larger and smaller models regarding the reduction of data available for fine-tuning in more detail. For this purpose, we utilize the F_1 scores shown in Figure 3 and correlate them with their model sizes noted in Table 1. We illustrate this in Figure 4 by considering two examples above and two examples below the mentioned threshold of 800 examples per category.

Model Groups   We see three groups of models. The previously mentioned group of large models (vit-base, vit-large, and beit-large), a group of medium-sized models (deit-base, beit-base, and focalnet-base), and finally the group of smaller models, among which we particularly want to monitor the subgroup of the smallest models (deit-tiny, resnet-18, and dinov2-small).

Larger Models Strengths   When examining the two representations in Figure 3 with over 800 examples per category, within the mentioned area of stable performance, two key observations emerge. Firstly, all model groups form performance clusters that are closely spaced, which means that their F_1 scores differ by less than 0.02 points. Secondly, there exists a noticeable trend: the larger the model, the better its performance. This observation aligns with the earlier conclusion that the group of larger models tends to dominate within the realm of stable performance.

Small Model Stability   However, when we drop below the mentioned threshold of 800 samples per category, three effects become evident. First, the performance clusters of the three model groups expand, i.e. the variance of the F_1 scores within the groups significantly increases. Second, the trend that larger models yield better results dissipates. The

most remarkable, however, is the third effect. Within the group of smaller models, which shows the most increase in variance, the performance of the subgroup of the smallest models remains surprisingly stable. This includes not only the two ViTs, deit-tiny and dinov2-small, but also resnet-18, the smallest CNN in our selection. The F_1 score of the ViT model dinov2-small with 200 examples per category is less than 0.01 points below that of the best large model, vit-large, with 1600 examples per category.

4.3. Practical Application Insights

Accuracy Evaluation   As outlined in Section 1, the original motivation of our work is to provide an automated quality control system using AI, even with a small number of examples. In practical application, particularly, the accuracy of the prediction of the fine-tuned models is a critical criterion for deployment. We have calculated this based on the test dataset introduced in Section 3 for the respective categories in Table 4. Motivated by previous results, we concentrated on the subgroup of the three smallest models as detailed in Table 1. These models were fine-tuned using example data from 200 examples per category.

Comparison   To contextualize the quality of these fine-tuned models in relation to existing solutions, results from a comparable study were considered, which were computed with Deep Convolutional Generative Adversarial Network (DCGAN) and Geometrical Transformation (GT) techniques for the CRFP use-case on a different dataset using synthetic image data augmentation (Meister et al., 2021). The results are not directly comparable since they were obtained on a different dataset. However, this allows for a rough assessment of the models calculated here. The nominal and gap categories are predicted with higher accuracy by all models presented here. For overlap, at least resnet-18 surpasses previous accuracies.

Table 4: Predicted accuracies on test dataset for categories nominal, overlap and gap for the group of smallest models in table 1.

Short Name	Nominal	Overlap	Gap
dinov2-small	95.72%	81.15%	96.78%
deit-tiny	92.64%	85.76%	97.80%
resnet-18	92.36%	90.27%	96.62%
DCGAN+GT[2]	88.37%	87.23%	93.41%

---

[2]Accuracies computed with Deep Convolutional Generative Adversarial Network and Geometrical Transformation techniques for CRFP use-case on a different dataset using synthetic image data augmentation (Meister et al., 2021).

(a) 200 examples per category.

(b) 400 examples per category.

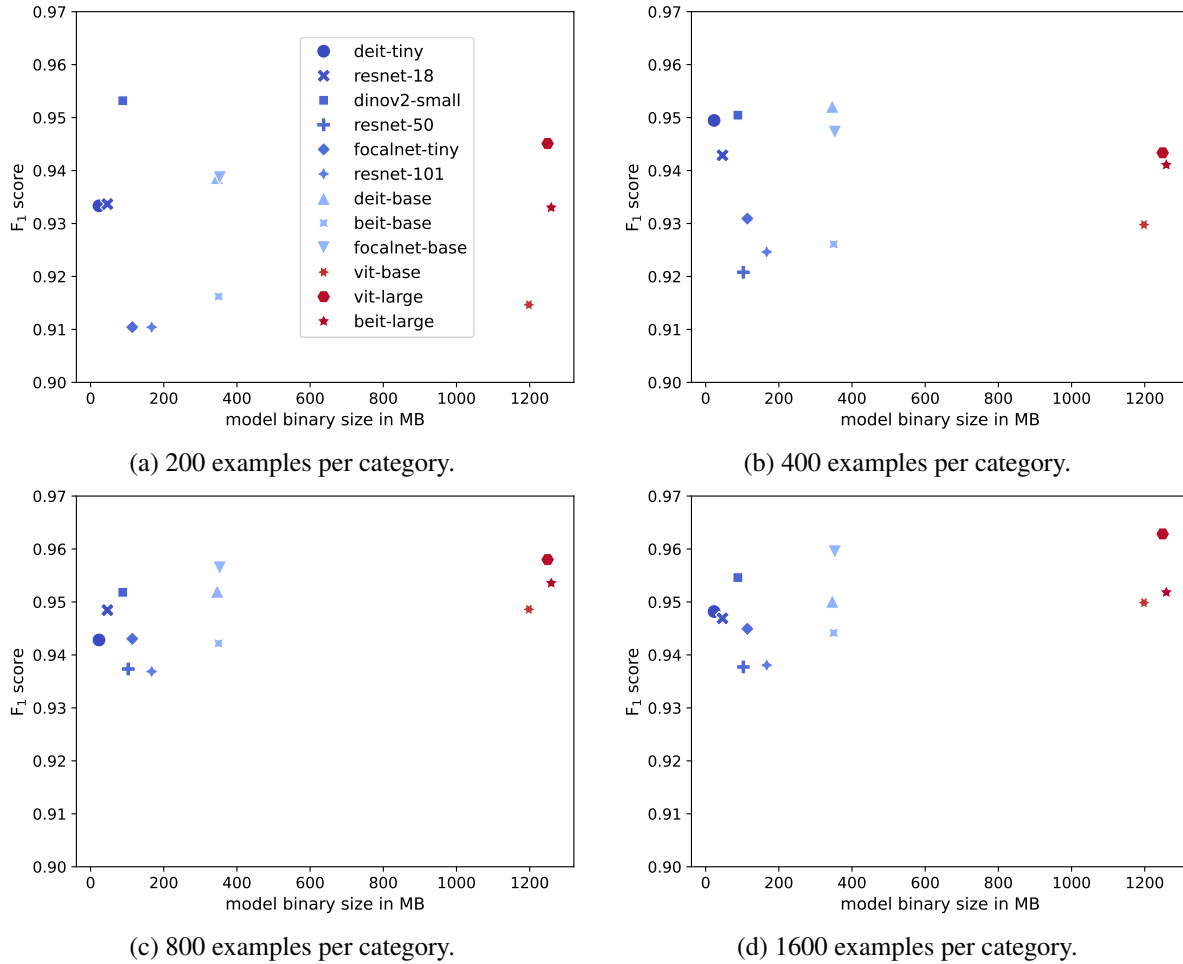(c) 800 examples per category.

(d) 1600 examples per category.

Figure 4: Analysis of performance relative to model size: Averaged F_1 scores versus PyTorch model binary file sizes for fine-tuned models from Table 1 across four training datasets with diverse category sample sizes.

5. Conclusion

Our investigation supports a comprehensive understanding of performance dynamics across different model architectures and sizes, which highlights the applicability of small models to small datasets and establishes the central role of transfer learning for our use case. Here we summarize our key findings:

Model Architectures Our study shows that ViTs and FMNs generally perform better than Convolutional Neural Networks (CNNs) on a range of datasets sizes. This supports previous research and introduces FMNs as a strong option. Since predicting the best model for a given task is difficult, incorporating a variety of models, such as ViTs and FMNs, in the search for the best hyperparameters is a practical approach.

Model Sizes Contrary to the common assumption that larger models generally perform better, our research shows that small models exhibit remarkable efficiency, especially when applied to smaller datasets. If this finding is transferable to other applications, it is recommended to experiment with a number of smaller models rather than a few large models when optimizing hyperparameters. This strategy offers two advantages: Saving computational resources and accelerating insights. In addition, the near-equivalent performance of small models, even with only 200 examples, compared to larger models trained with 1600 examples, offers significant advantages for edge computing devices. By enabling local high-quality predictions, small models reduce bandwidth requirements and ensure low latency in fast production environments. In addition, the ability to specialize and combine multiple small models for different fault classifications opens up new ways to achieve exceptional accuracy with minimal resources.

Use-Case Application The results presented in section 4 confirm the competitiveness, perhaps even the superiority, of transfer learning as a method in contexts constrained by limited data. The fine-tuned models not only demonstrate the feasibility of productive use, but also offer a compelling alternative to manually adapting algorithms. This research shows that with transfer learning, we are able to develop efficient AI models tailored to specific application requirements even with limited data.

Our research supports a paradigm shift towards resource-efficient, adaptable and accessible AI model development. By leveraging the strengths of Open Source ViTs, FMNs and the strategic use of transfer learning, we are demonstrating the way for innovative solutions that are not only computationally efficient, but also offer robust performance in a variety of use cases.

Impact Statement

This paper presents work aimed at advancing the field of Machine Learning and CFRP Tape Laying. There are many potential technical and economic consequences of our work in industrial production, but none that we feel must be specifically highlighted here.

References

Bao, H., Dong, L., and Wei, F. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. URL https://arxiv.org/abs/2106.08254.

Cireşan, D., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. High-performance neural networks for visual object classification. *Computing Research Repository - CORR*, 02 2011.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. *Neural Computation*, 22 (12):3207–3220, 12 2010. ISSN 0899-7667. doi: 10.1162/NECO_a_00052. URL https://doi.org/10.1162/NECO_a_00052.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., Gelly, S., Houlsby, N., Zhai, X., and Lucic, M. On robustness and transferability of convolutional neural networks. *CoRR*, abs/2007.08558, 2020. URL https://arxiv.org/abs/2007.08558.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36 (4):193–202, Apr 1980. ISSN 1432-0770. doi: 10.1007/BF00344251. URL https://doi.org/10.1007/BF00344251.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hubel, D. H. and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962. doi: https://doi.org/10.1113/jphysiol.1962.sp006837. URL https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1962.sp006837.

Hugging Face. Model Hub, 2024. URL https://huggingface.co/models.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Large scale learning of general visual representations for transfer. *CoRR*, abs/1912.11370, 2019. URL http://arxiv.org/abs/1912.11370.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Li, L., Jamieson, K. G., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., and Talwalkar, A. Massively parallel hyperparameter tuning. *CoRR*, abs/1810.05934, 2018. URL http://arxiv.org/abs/1810.05934.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. Tune: A research platform for distributed model selection and training. *ArXiv*, abs/1807.05118, 2018. URL https://api.semanticscholar.org/CorpusID:49741140.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Meister, S., Möller, N., Stüve, J., and Groves, R. M. Synthetic image data augmentation for fibre layup inspection processes: Techniques to enhance the data set. *Journal of Intelligent Manufacturing*, 32(6):1767–1789, Aug 2021. ISSN 1572-8145. doi: 10.1007/s10845-021-01738-7. URL https://doi.org/10.1007/s10845-021-01738-7.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

Schuster, A., Mayer, M., Willmeroth, M., Brandt, L., and Kupke, M. Inline quality control for thermoplastic automated fibre placement. *Procedia Manufacturing*, 51:505–511, 2020. ISSN 2351-9789. doi: https://doi.org/10.1016/j.promfg.2020.10.071. URL https://www.sciencedirect.com/science/article/pii/S2351978920319284. 30th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2021).

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, pp. 1–14. Computational and Biological Learning Society, 2015.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19175–19186, June 2023.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Yang, J., Li, C., Dai, X., and Gao, J. Focal modulation networks. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ePhEbo039l.

Zhou, H.-Y., Lu, C., Yang, S., and Yu, Y. Convnets vs. transformers: Whose visual representations are more transferable? In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2230–2238, 2021. doi: 10.1109/ICCVW54120.2021.00252.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. V. Rethinking pre-training and self-training. NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

A. Code and Data Accessibility

For the purpose of review and to facilitate reproducibility of our findings, we have provided an anonymized version of our code repository. It can be accessed through the following link:

https://anonymous.4open.science/r/crfp-transfer-learning/README.md

We encourage reviewers and interested parties to consult the README file within the repository for detailed instructions on how to access and utilize the code. For accessing the necessary dataset, please use the provided S3 credentials:

```
AWS_ACCESS_KEY_ID="GOOG1E7FETDFKZD3RQB75RPGQAXOELXNY4QEKLFO3J3BM2NTVGO4HSROJ24IA"
AWS_SECRET_ACCESS_KEY="PhQK7ymyI6UDaxeytOyWr+AiHCQOnDKgFl/5KMTH"
```

It is our intention to publish the code repository and make the selected models and data publicly available on GitHub and Hugging Face, respectively, upon conclusion of the review process. This will ensure that our contributions to the field are accessible for further research and application in the broader AI and machine learning community.