2024 EC³
2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 15-17, 2024

A COMPARATIVE STUDY OF DEEP LEARNING MODELS FOR GRANULOMETRY IMAGE BASED ESTIMATION OF CONCRETE AGGREGATE

Benjamin Pasquier and Houda Chabbi Drissi
iCoSys - Institute of Artificial Intelligence and Complex Systems
HEIA-FR, HES-SO Haute école spécialisée de Suisse occidentale

Abstract

Obtaining the granulometry is the starting point of our pipeline for automating the calculation of concrete properties using images. For this reason, we focus on developing the best deep learning model that can compute aggregate gradation and can generalize to images obtained from different aggregate producers. We investigate two established approaches: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Our analysis includes a dedicated CNN model trained from scratch, alongside pre-trained CNN and ViT models adapted through transfer learning.

To evaluate the performances and the generalization ability of the models, we use three different datasets: two publicly available and one of our own. Our analysis shows that transfer learning followed by fine-tuning on ViT_16 outperforms the other models, on both classification and regression tasks, with smaller errors and greater generalization capabilities.

Introduction

Deep learning applied to images is successfully used for prediction in many fields. Our aim is to use images of aggregates and a deep learning model to predict the properties of concrete. Our assumption is that such a model, trained on a dataset of aggregate images and corresponding concrete property data, would extract meaningful features from the visual representations of aggregates. These features can then be used to predict concrete properties such as compressive strength and workability.

The most influential factor on the properties of concrete mixtures is the type of used aggregates and their granulometry. Therefore, we propose to build a system that uses aggregate images as the foundation for the predictive final system, which will then use additional inputs to predict concrete properties based on the aggregate size distribution, as shown in Figure 1. Integrating this model with a camera system observing the aggregate conveyor belt during concrete production allows for real-time granulometry determination through image analysis. Leveraging this real-time information, the entire system can then continuously estimate key concrete properties enabling real-time monitoring of the mix design and ensuring consistent concrete quality.

We are interested in concrete mixes in which recycled or natural aggregates may be present in varying proportions. Therefore, we seek for a model that can accurately and automatically extract granulometry distribution from aggregates images. In addition, this model should have a strong generalization capability, adapting to new images and new particle size distributions of natural or recycled aggregates, without the need to explicitly specify the aggregate type.

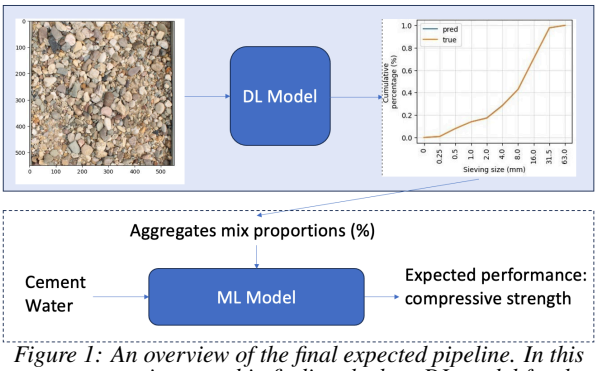

Figure 1: An overview of the final expected pipeline. In this paper, we are interested in finding the best DL model for the granulometry task.

To carry out our study, we compare the performances of different deep learning architectures for granulometry estimation, first as a classification task, then as a regression task. To obtain a deep learning model, two main approaches can be considered: either build an own network and train it, or take advantage of generalist pre-trained models to perform the desired downstream task using transfer learning. The latter leverages knowledge from large datasets to improve efficiency and performance on new tasks. In our evaluation, we take AggNet, a specialized Convolutional Neural Network (CNN) model (Coenen et al., 2022) on the one hand, and perform transfer learning on pre-trained CNN models from different families on the other: MobileNetV2 (Sandler et al., 2018), ResNet50 (He et al., 2015) and finally the Vision Transformer (ViT) model ViT_16 (Dosovitskiy et al., 2020), a recent architecture using self-attention mechanisms (Vaswani et al., 2017).

In addition to leveraging transfer learning to enhance the performance of pre-trained models, we use hyperparameter optimization techniques to refine the predictions of these models. ResNet50 and MobileNetV2 are two popular deep learning CNN models that has been shown to be

effective for a wide range of tasks and are relatively easy to train and fine-tune. ViT_16 is a vision transformer (ViT) that also shown to be state-of-the-art for image classification. ViT architecture models are interesting in this study because they are expected to allow a better accuracy and generalize better than CNNs (Maurício et al., 2023).

The paper is organized as follow: the next section reviews related works. Then, the two following sections introduce our methodology and our experimental setup to carry on the study for classification and regression tasks. Before concluding, we present the results of our experiments.

Related Work

Classification and granulometry tasks to determine the distribution of particle is very important not only for estimating concrete properties but for ore field in general. To avoid the need for costly manual techniques such as sieving, work has been carried out to automate the process by analyzing images of aggregates or ores. Since the emergence of deep learning, research has embraced CNN as a primary approach. A comprehensive survey on ore image processing using deep learning can be found in WANG Wei and Hao (2023), which highlights similarities to approaches employed for aggregates. The application of deep learning for aggregates can be divided into two main categories: those that start from scratch by building their own models, as seen in Lau Hiu Hoong et al. (2020), Qin et al. (2023) or Coenen et al. (2022), and those that leverage existing pre-trained models and employ transfer learning techniques to adapt them to specific tasks, such as Olivier et al. (2020). In addition to this distinction, there is another categorization based on the image processing strategy employed. The first one is to classify individual aggregate images as described in Sun et al. (2022), while the second way is to regress from images of aggregate mixtures to granulometry distributions.

In the first category, the authors of Lau Hiu Hoong et al. (2020) propose a customized Residual Network (ResNet) model and a dataset of 36'000 images of individual grains. They achieve a classification accuracy of 97% (brick, ceramic, stone, etc.). They also proposed a segmentation method to predict the nature of each grain in an image of a multi-grain sample. In Qin et al. (2023), the study is based on the concept of instance segmentation, using a specialized neural network model (AS Mask RCNN) to detect and classify individual aggregates within mixed aggregate images. The results of their study indicate that the AS Mask RCNN model achieved an accuracy of over 89.13%. The approach requires a dataset made up of images and their segmentation masks for each aggregate to be provided for each image. All these papers demonstrate the relevance of using deep learning models when calculating granulometry based on the segmentation of individual aggregates.

In the second category, where classifying and regressing are used without segmentation, deep learning has also been successfully used. In Olivier et al. (2020), the authors use the CNN architecture VGG16 Simonyan and Zisserman (2015) with transfer learning to predict the ten size fractions considered for an ore. The obtained results show the effectiveness of a CNN in predicting the size distribution of ore, with a mean model error of -0.012 and a standard deviation of 0.107. Coenen et al. (2022) presents a deep learning model, AggNet, for real-time determination of concrete aggregate grading curves. They propose a dedicated CNN network model with multi-scale feature extraction to handle diverse particle sizes and showed good results on a classification task with an accuracy of 95.5%, which is the best according to our knowledge.

We are interested in the second category of approaches because of their simplicity and industrial applicability. Once trained, these models dispense with the need for labor-intensive data preparation, allowing for direct estimation of aggregate granulometry from images. The main difficulty lies in preparing a dataset of varied aggregates images with their size distribution. Authors of AggNet published their dataset Coenen (2022) which we rely on as it meets our needs: each aggregate image is associated with its particle size distribution. We use this model as a reference for our analysis to study how it generalizes to our own dataset and to compare it to our approach which is based on adapting pre-trained models. In Coenen et al. (2023), the authors propose to use vision transformers and developed again their own model based on this architecture. They demonstrate the technical feasibility and interest of this approach. However, we believe that we can leverage the feature extraction capabilities of the pre-trained models (CNN or ViT), acquired through training on vast image datasets, and tailor them to our downstream tasks of classification and regression.

Methodology

As said in the previous sections, we compare four neural network architectures for estimating the granulometry of aggregates from images. We separate this task into two sub-tasks: the first one aim to classify the aggregates images toward the corresponding DIN 1045-2 Deutsches Institut für Normung (2008) standard granulometry class and the second one aim to directly estimate the mass percentage for each bin size considered. To perform this comparative study, we use three different datasets, two are publicly available and one is own-made.

Data

We use two publicly available datasets that contain image samples of natural aggregates: the Visual Granulometry dataset (Coenen, 2022) and the Deep Granulometry dataset (Coenen, 2023). The first one is designed for a classification task and the latter for a regression task. The Visual Granulometry dataset contains 900 images of aggregates along with their corresponding DIN 1045-2 standard granulometry class. There are nine classes in the standard (see Figure 2), each representing a grading curve, i.e. the size distribution of the aggregates. For each class, two samples of 5 kg of aggregates were produced and mixed

to obtain a total of 50 images per sample, and 100 images per class.

The Deep Granulometry dataset contains 1650 images of coarse aggregate samples with different particles sizes ranging from 0.1 mm to 32 mm. Each image is accompanied by the mass percentage of each particle size bin considered, following 33 different granulometries (11 per largest grain size).

We then use a custom dataset with our own data that we use only for evaluation purposes, in order to measure the generalization of the trained models. This dataset contains 174 images of both recycled or natural aggregates from seven different sources, i.e. seven different granulometries, in an unbalanced fashion. For the classification task, we assign the DIN 1045-2 class to each of these granulometries by minimizing the mean squared error between them and the granulometry of each class. As shown in Figure 2, our grading curves can be far from the standard classes, therefore the assignation is not exact but still allows us to evaluate the model on our own data for a classification task. As two granulometries fall in the same class, we have only 6 of the 9 DIN 1045-2 classes that are represented. For the regression task, we simply report the mass-percentage for each size bin considered in the Deep Granulometry dataset.

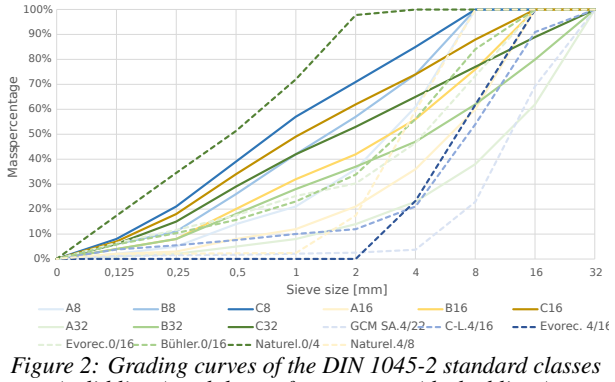

Figure 2: Grading curves of the DIN 1045-2 standard classes (solid lines) and those of our sources (dashed lines).

Since the two publicly available datasets contains images rectified by homography that were taken with a ground sampling distance (GSD) of 0.125mm, we also rectified our images in a similar way. However, as the setup was not exactly the same for all taken images, 80 of images have been cropped manually and can therefore present some deformations due to the perspective. In addition, the GSD of our images is not exactly equal to 0.125mm. These variations in our own dataset will serve to assess the generalization capabilities of the models we test. The Table 1 summarizes the size and characteristics of the two datasets used.

Neural networks

Classification, i.e. classifying images of aggregates towards the right DIN1045-2 standard class, is the first task we consider. We assume that this task is simpler than the regression one which consists of predicting the real percentage for each size bin considered. Therefore, we evaluate three CNN-based models, namely ResNet, MobileNetV2 and AggNet, as well as one ViT model. The AggNet model is a dedicated CNN model for granulometry estimation and the source code of its architecture has been made available by the authors (Coenen, 2022). The remaining three models are pre-trained models that we adapt to the granulometry estimation task using transfer learning. We then adapt and evaluate the two best performing classification models on the regression task, i.e. estimating the mass percentage for each size bin considered.

Transfer learning

As the state-of-the-art computer vision models are composed of millions of parameters, they need to be trained on large datasets. In order to adapt these models on a new task, it is often recommended to use the weights of a pre-trained model instead of training the model from scratch and risking to overfit the data. This can be done by freezing the weights of the pre-trained model and adding a new fully connected layer on top of it, which will be trained on the new task. This process is called transfer learning. In this study, we freeze the weights of the pre-trained model feature extraction layers, using it as a feature extractor, and train a new fully connected layer on top of them, as shown in Figure 3.

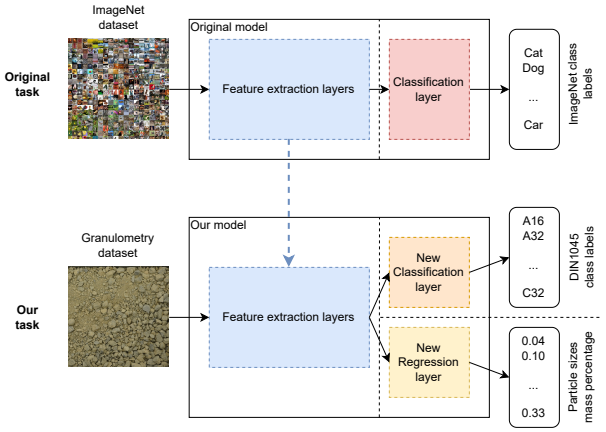

Figure 3: An overview of the transfer learning process used in this study. Feature extraction layers of the pre-trained model are frozen and new trainable classification or regression layers are added on top of them. Top image from (Deng et al., 2009).

To perform our comparative study, we use transfer learning on three pre-trained models, namely ResNet50, MobileNetV2 and ViT_16, all originally trained on the ImageNet (Deng et al., 2009) dataset.

Training

The network parameters θ are learned by optimizing a loss function $\mathscr{L}(\theta)$, which differs depending on the network task. The loss function is computed for each batch of data and the network weights are updated according to the gradient of the loss function with respect to the weights. The compared classification networks aim to classify images of aggregates towards M classes, where M is the number of DIN 1045-2 standard classes. To do so, they are

Table 1: Summary of the datasets used.

Dataset	Task	Size	Images size	Particles size	Nb. of classes / size bins
Visual Granulometry (Coenen, 2022)	Classification	900	2200x3000px	0-32mm	9
Deep Granulometry (Coenen, 2023)	Regression	1666	2200x3000px	0-32mm	33
Own	Both	174	1072x1472px	0-32mm	6 / 7

trained by minimizing the well-known cross-entropy loss function, defined as

$$\mathscr{L}_{CE}(\theta) = -\sum_{i=1}^{M} y_i \log(\hat{y}_i) \quad (1)$$

where y_i is the ground truth label for the i^{th} class (either 0 or 1) and $\hat{y}_i$ is the predicted probability for the i^{th} class.
For the regression task, we aim to predict the mass percentage of the M size bins considered. Therefore, the networks are trained by minimizing the Kullback-Leibler divergence, which is the same loss used in (Coenen et al., 2022). It is defined as

$$\mathscr{L}_{KL}(\theta) = \sum_{i=1}^{M} y_i \log(\frac{y_i}{\hat{y}_i}) \quad (2)$$

where y_i is the ground truth mass percentage for the i^{th} size bin and $\hat{y}_i$ is the predicted mass percentage for the i^{th} size bin.
For training, we systematically perform early stopping to avoid overfitting and use the Adam optimizer (Kingma and Ba, 2017) to optimize model weights. In order to obtain the best performances, we then perform different optimizations and evaluate their impact on a validation set before selecting the best model.
– **Data augmentation :** As the dataset used for both classification and regression are relatively small (900 to 1666 images), we perform data augmentation to allow models to generalize better and avoid overfitting on training data. This data augmentation is performed during the training, on the fly, so that it is highly unlikely for the model to see the same image twice. We test two kind of data augmentation:

1. A tuned augmentation by evaluating the model many times on different combinations of geometric transformations, such as rotation, shift, zoom or shear.

2. The augmentation proposed in paper (Coenen, 2022), performing geometric and radiometric (e.g. hue shift) transformations.

We compare the best tuned augmentation to the one proposed in (Coenen, 2022) and keep the one that performs the best on the validation set.
– **Hyperparameter tuning :** We then tune the hyperparameters of the models in order to optimize their performances. This is done by evaluating many times the model on different combinations of hyperparameters, which are the batch size, the neurons number in the fully connected layers, the dropout rate and the learning rate. We keep the combination of hyperparameters that gives the best performances on the validation set.
– **Fine-tuning :** Finally, we perform a final fine-tuning by unfreezing the weights of the pre-trained model feature extraction layers and continuing the training for a few epochs. We then evaluate if this fine-tuning improves the performances of the model on the validation set.

Experimental Setup

In order to evaluate the performances of the different models, we use an identical experimental setup for all of them. We first split the dataset into a training and test set with a ratio of 80% and 20% respectively, the latter being only used for the final evaluation of each model. Before training, the training set is further split into a training and validation set with the same ratio, allowing us to select the best model according to different configurations and to ensure its good generalization. The Table 2 summarizes the number of images used for training and evaluation for each task.

Table 2: Number of images used for training and evaluation for each task. Evaluation is both performed on the Deep or Visual Granulometry (V/DG) datasets and on our own dataset.

Set		Classification	Regression
Training		720	1326
Evaluation	V/DG	179	340
	Own	174	174

As the architecture of the three pre-trained models we consider are designed to take images of 224x224 pixels as input, all the images are cropped to be squared and then resized to this size. For the AggNet model, images are simply down-sampled to 550x750 pixels, keeping their original size ratio.
Figure 4 summarizes the training and evaluation procedures. More details for each task we consider are given in the next sections.

Classification task

For the classification task, we train and evaluate respectively four models : ResNet50, MobileNetV2, ViT_16 and AggNet. Hyperparameter tuning, data augmentation tuning and fine-tuning are only performed on ResNet50 and MobileNetV2, as ViT_16 showed already good performances without doing so (see section *Results and dis-*

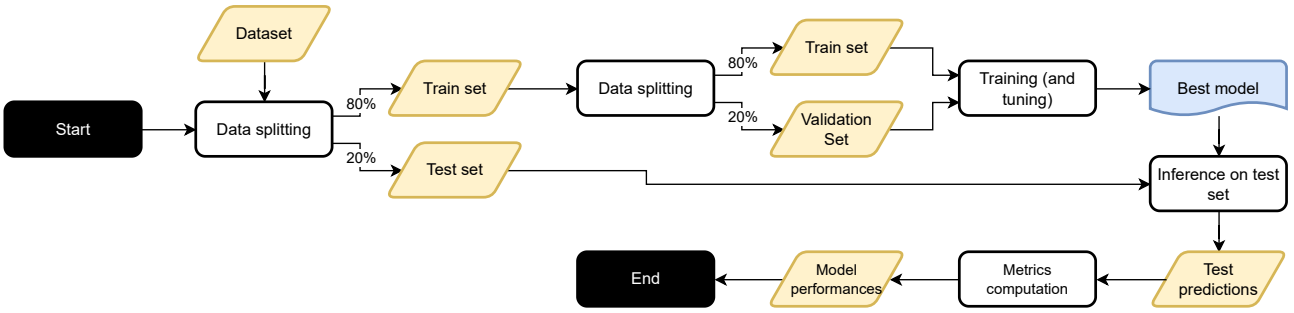

Figure 4: Training and evaluation procedure for each model architecture we compare, on both classification and regression tasks.

cussion) and as the authors of the AggNet model already performed these optimizations.
To evaluate their performances, we use the accuracy metric, defined as

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (3)$$

We also compute the confusion matrix, that allows us to see which classes the evaluated model confuses the most. It also allows us to quickly calculate other metrics per class, such as the precision, recall and F1-score.

Regression task

For the regression task, we only train the two best performing models on the classification task, namely ViT_16 and AggNet (see section *Results and discussion*). This time, we also perform hyperparameter tuning, data augmentation tuning and fine-tuning on ViT_16 in order to obtain the best possible model and see if it can outperform AggNet. Training configuration for AggNet model is once again taken from (Coenen, 2022), while the best configuration we find for ViT is the following :

- **Tuned hyperparameters** : batch size of 64, dropout rate of 0, hidden size of 512 and learning rate of 0.007
- **Tuned data augmentation** : horizontal and vertical flip, shift range of 0.1, zoom range of 0.3 and fill mode on reflect.
- **Fine-tuning** : learning rate of 0.0001.

To evaluate the performances of each model, we use the mean absolute error (MAE) metric and the root mean squared error (RMSE) metric, defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

and

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (5)$$

where N is the number of predictions, y_i is the ground truth mass percentage for the i^{th} image and $\hat{y}_i$ is the predicted mass percentage for the i^{th} image. These metrics are calculated for each size bin and then averaged to obtain a single measure of model performances.
We compute the RMSE along with the MAE because the RMSE penalizes more the large errors than the MAE, and thus gives us an other important information about the model performances. We use the RMSE instead of the MSE because it is more interpretable as it is in the same unit as the ground truth mass percentage vector.

Results and discussion

We first discuss the results obtained by the different classification models, before focusing on the results obtained by the two best performing ones on the regression task.

Classification

Table 3 summarizes the performances in terms of accuracy obtained by each compared models on the two test datasets. While they all performs significantly better on the Visual Granulometry dataset, the ViT_16 outperforms the other models with respective accuracies of **97%** and **34%** on both datasets. As neither hyperparameter tuning nor data augmentation tuning were performed on ViT_16, we can assume that even better performances could be obtained by doing so. This shows how powerful transformers can be on various tasks, including computer vision tasks. The AggNet model is also performing very well on the Visual Granulometry dataset Coenen (2022), which coincide with the results reported in Coenen et al. (2022). All the models show a low accuracy on our own dataset, with accuracies ranging from 26% (AggNet) and 34% (ViT_16), showing a poor generalization of the model on the classification task.

Table 3: Performances of the different classification models on the two test sets, i.e. the Visual Granulometry (VG) and our own dataset.

Model	Accuracy	
	VG data	Our data
ResNet-50	0.85	0.30
MobileNetV2	0.87	0.29
ViT_16	**0.97**	**0.34**
AggNet	0.94	0.26

Figure 5 shows the confusion matrix obtained by the ViT_16 model on our own dataset and help to understand why models are under performing on it. As it shows high

Table 4: Regression results on the Deep Granulometry dataset with both models. Errors are computed for each grain size bins considered and then averaged.

Grain size bins [mm]		**0.25**	**0.5**	**1**	**2**	**4**	**8**	**16**	**31.5**	**63**	**Avg.**
AggNet	MAE [%]	0.22	1.23	1.36	0.67	1.65	1.61	1.73	1.69	0.21	1.15
	RMSE [%]	0.28	1.56	1.78	0.86	2.11	2.18	2.61	2.72	0.47	1.62
ViT	MAE [%]	0.13	0.88	0.86	0.46	0.51	0.98	0.94	0.49	0.04	**0.59**
	RMSE [%]	0.18	1.20	1.29	0.66	0.73	1.59	1.61	1.02	0.14	**0.93**

accuracy on Visual Granulometry data, we do not present here the confusion matrix obtained on this data. On our own data, ViT_16 frequently misclassifies B16 as A32 or A16, and A32 as A16, likely due to the fact that our data samples do not exactly follow the particle size distribution of the classes defined by DIN 1045. For better performances, we should add more classes instead of simply assign our sample to one of the standard classes. Analyzing ViT's regression performances on our data might be more revealing, as it predicts continuous values of the real granulometry instead of inferred discrete classes.

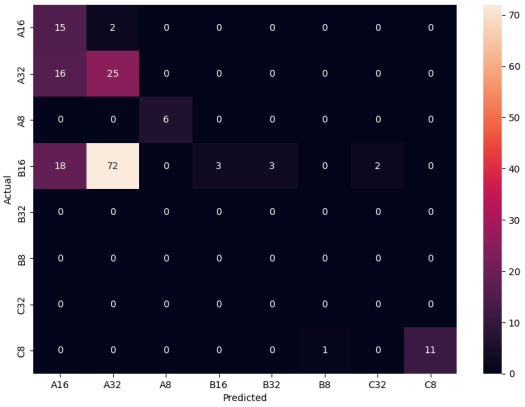

Figure 5: Confusion matrix obtained on our own dataset with the ViT_16 model.

Regression

Since the best models on the classification task on Visual Granulometry data are the ViT_16 and AggNet models, we only train and evaluate these for a regression task.

Table 4 first shows the results obtained by both models on the Deep Granulometry dataset (Coenen, 2023). The ViT model therefore fares better than the AggNet model, with an average MAE of **0.59%** versus 1.15%, i.e. half as much. The obtained RMSE with both models also confirmed this observation. As far as errors by size are concerned, the ViT model seems to have more difficulty in predicting proportions for sizes from 0.5mm to 1mm and from 4 to 31.5mm, as does the AggNet model. It shows that either some sizes are more difficult to differentiate than others, either the data is much more varied in certain bin sizes than others.

The worst respectively the best predicted grading curves by ViT on this first dataset are shown in Figure 6, along with the ground truth. We see that the model achieve to predict the perfect grading curve in some case, and to predict a grading curve that is still very close from the ground truth in the worst case.

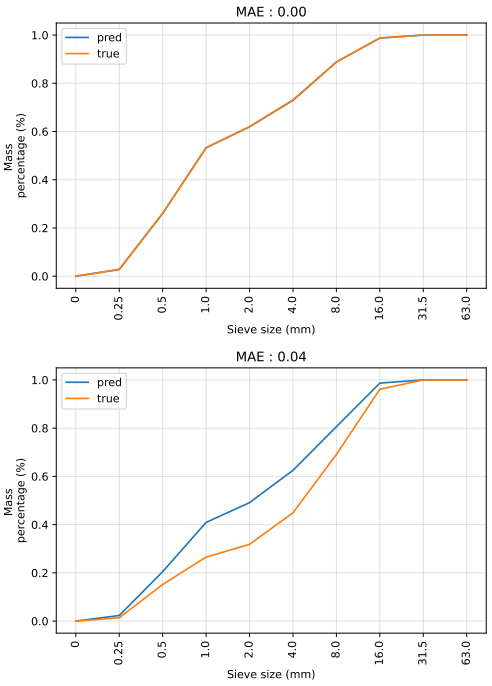

Figure 6: Best (top) and worst (bottom) predicted grading curves (in blue) by ViT on the Deep Granulometry dataset, along with the ground truth (in orange)

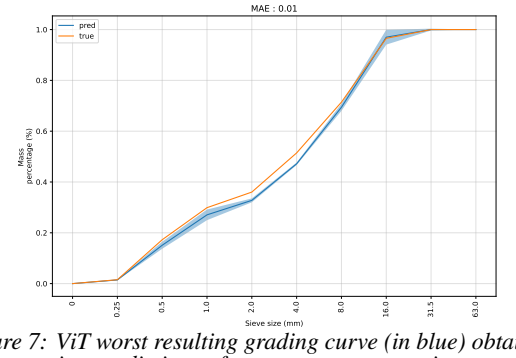

Figure 7: ViT worst resulting grading curve (in blue) obtained by averaging predictions of a same aggregates mixture on the Deep Granulometry dataset, along with the ground truth.

If we average predictions over samples following the same grading curve, we obtain a new prediction that is much closer to the ground truth even in the worst case, as shown in Figure 7. This shows that in a real setup where many images of the same mixture of aggregates are taken, we can predict a much more precise granulometry by averaging the predictions made by the model.

Table 5 then shows the results obtained by both models on our own dataset. The results are significantly worse than those obtained on the Deep Granulometry dataset, with an

Table 5: Regression results on our own dataset with both models. Errors are computed for each grain size bins considered and then averaged.

Grain size bins [mm]		**0.25**	**0.5**	**1**	**2**	**4**	**8**	**16**	**31.5**	**63**	**Avg.**
AggNet	MAE [%]	6.06	9.0	7.22	4.64	9.46	9.87	9.04	13.33	0.41	7.67
	RMSE [%]	10.01	9.45	7.88	5.44	10.15	12.07	11.66	15.63	0.71	9.22
ViT	MAE [%]	6.35	4.57	4.14	5.00	8.47	11.87	9.53	10.82	0.18	**6.77**
	RMSE [%]	10.22	5.28	4.98	6.21	10.76	13.82	11.44	12.97	0.29	**8.44**

average MAE of **6.77%** for the ViT model and 7.67% for the AggNet model. While these errors are similar, ViT is still able to generalize better that the AggNet model. This overall increase in errors can be explained by the difference in particle size distributions between the two datasets. Indeed, the granulometries of our dataset are significantly different from those of the Deep Granulometry dataset. As the models are trained on the latter, it is very likely that they will have difficulty generalizing to other data. In addition, some images in our dataset have not been rectified by homography, this difference in images may therefore also have an influence on model performances.

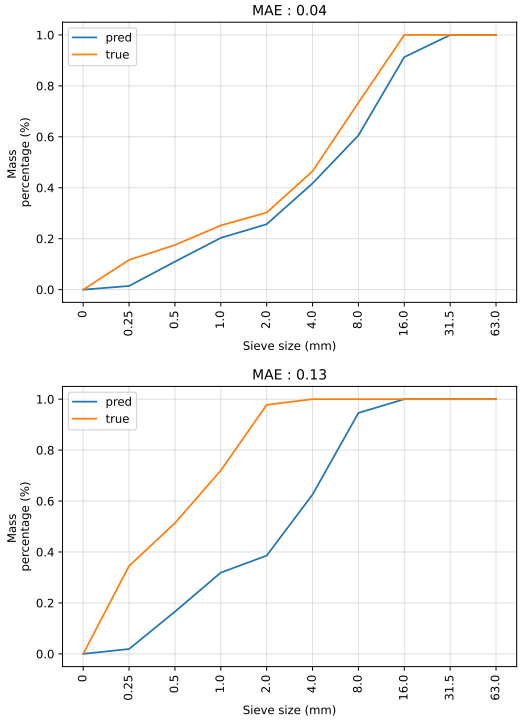

Figure 8: Best (top) and worst (bottom) predicted grading curves (in blue) by ViT on the Deep Granulometry dataset, along with the ground truth (in orange)

The worst respectively the best predicted grading curves by ViT on our dataset are shown in Figure 8, along with the ground truth. This time, we see that the model may struggled to predict a grading curve close from the ground truth, especially when aggregates follow a granulometry far from the ones the model was trained on. These results indicate that we may need aggregates training data that follows a wider range of granulometries in order to increase the generalization of the model. We can still note that for different but close granulometries, the model is able to make prediction with few errors, which is encouraging. If we average predictions over samples following the same grading curve, we achieve to reduce the MAE but still remains relatively high (11%) in the worst case, as shown in Figure 9. The need for more varied training data therefore remains.

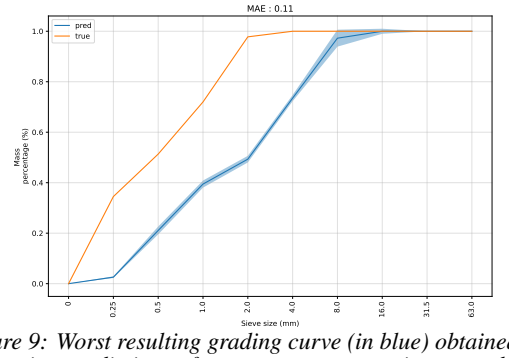

Figure 9: Worst resulting grading curve (in blue) obtained by averaging predictions of a same aggregates mixture made by ViT on our own dataset, along with the ground truth (in orange).

Conclusion

The result of our study shows that using the dedicated CNN AggNet model of Coenen et al. (2022) for the task of classification and calculation of the distribution of aggregates from their images is a better approach than performing transfer learning on the most popular pre-trained CNNs that we used: ResNet50 and MobileNetV2. On the other hand, our tests show that applying transfer learning on a pre-trained model based on transformers (ViT_16) allows to achieve better results on the two considered tasks. Iman et al. (2023) positions transfer learning as a valuable technique to unlock the full potential of deep learning. In our case, where the images are very specific and different from datasets of the pre-trained models, the smaller AggNet model, using multi-scale feature extraction layers, effectively handles the diverse aggregate sizes compared to the generalist pre-trained CNNs. This specialization likely contributes to its better performance. Similarly, ViT_16 excels in this task due to the inherent ability of transformers to capture both local and global interactions within the image, potentially explaining the advantage of ViT over AggNet.

To evaluate the generalization ability of each model, we employed our own dataset for testing. This dataset differs from the publicly available one used to train and evaluate the models in two key ways. Unlike the public dataset, it includes both natural and recycled aggregates and it presents different aggregate size distributions. Again, we obtained better results with ViT_16, which reinforces the

idea of deepening this approach to improve its adaptability to datasets that do not perfectly follow the granulometry of the training data. While the current results of this generalization are not optimal, we believe incorporating a subset of our own data into the training set has the potential to significantly improve model performance.
Therefore, our next future task is to enlarge our own dataset in order to cover a greater variety of grading curves. Besides, our dataset is currently imbalanced, and we aim to achieve a balanced distribution with at least 100 images per class. While this imbalance was not a disadvantage for the present study (used for testing only), it needs to be addressed to determine the best strategy for ViT_16 generalization. We will explore two options: fine-tuning the model on a subset of our own data or reapplying transfer learning with hyperparameter tuning incorporating this subset into the training dataset. Our goal is also to identify the optimal dataset size that maximizes the generalization performance of the ViT model for granulometry estimation.

Acknowledgments

We are thankful to Daia Zwicky, Julien Ston and Sonia Anselmina from iTeC institute of HEIA-FR (Institut des Technologies de l'Environnement Construit) for the preparation of our own dataset.
This research was partially supported by a grant of the Programme de recherche HEIA-FR / Smart Living Lab (AGP:119149).

References

Coenen, M. (2022). Dataset: Visual granulometry: Image-based granulometry of concrete aggregate.

Coenen, M. (2023). Dataset: Deep granulometry.

Coenen, M., Beyer, D., and Haist, M. (2023). Granulometry transformer: image-based granulometry of concrete aggregate for an automated concrete production control.

Coenen, M., Beyer, D., Heipke, C., and Haist, M. (2022). Learning to sieve: Prediction of grading curves from images of concrete aggregate. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2022:227–235.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255.

Deutsches Institut für Normung, B. (2008). Din 1045-2: Concrete, reinforced and prestressed concrete structures.

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. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. CoRR, abs/2010.11929.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. CoRR, abs/1512.03385.

Iman, M., Arabnia, H. R., and Rasheed, K. (2023). A review of deep transfer learning and recent advancements. Technologies, 11(2).

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Lau Hiu Hoong, J. D., Lux, J., Mahieux, P.-Y., Turcry, P., and Aït-Mokhtar, A. (2020). Determination of the composition of recycled aggregates using a deep learning-based image analysis. Automation in Construction, 116:103204.

Maurício, J., Domingues, I., and Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. Applied Sciences, 13(9).

Olivier, L. E., Maritz, M. G., and Craig, I. K. (2020). Estimating ore particle size distribution using a deep convolutional neural network this work is based on research supported in part by the national research foundation of south africa (grant number 111741). IFAC-PapersOnLine, 53(2):12038–12043. 21st IFAC World Congress.

Qin, J., Wang, J., Lei, T., Sun, G., Yue, J., Wang, W., Chen, J., and Qian, G. (2023). Deep learning-based software and hardware framework for a noncontact inspection platform for aggregate grading. Measurement, 211:112634.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. CoRR, abs/1801.04381.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Sun, Z., Li, Y., Pei, L., Li, W., and Hao, X. (2022). Classification of coarse aggregate particle size based on deep residual network. Symmetry, 14(2).

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

WANG Wei, LI Qing, Z. D.-z. L. H. and Hao, W. (2023). A survey of ore image processing based on deep learning.