DEEP LEARNING-BASED ROCK PARTICULATE CLASSIFICATION USING ATTENTION-ENHANCED CONVNEXT

Anthony Amankwah [1] and Chris Aldrich [2]

[1]Amankwah Consult, Schwabhauserstr 3 82669 Geltendorf , Germany
[2]Curtin University, Faculty of Engineering, Sydney Australia

ABSTRACT

Accurate classification of rock sizes is a vital component in geotechnical engineering, mining, and resource management, where precise estimation influences operational efficiency and safety. In this paper, we propose an enhanced deep learning model based on the ConvNeXt architecture, augmented with both self-attention and channel attention mechanisms. Building upon the foundation of ConvNext, our proposed model, termed CNSCA, introduces self-attention to capture long-range spatial dependencies and channel attention to emphasize informative feature channels. This hybrid design enables the model to effectively capture both fine-grained local patterns and broader contextual relationships within rock imagery, leading to improved classification accuracy and robustness. We evaluate our model on a rock size classification dataset and compare it against three strong baseline. The results demonstrate that the incorporation of attention mechanisms significantly enhances the model's capability for fine-grained classification tasks involving natural textures like rocks.

KEYWORDS

Covnext, self-attention , channel attention

1. INTRODUCTION

Rock size classification plays a critical role in various industrial applications such as mining, blasting optimization aggregate quality control. Traditional manual and semi-automated methods are time-consuming and subject to human error, motivating the development of automated vision-based. The automated vision-based methods can be divided into classical image processing methods [9],[10] and machine learning/deep learning methods[7][8]. In this work, a deep learning method is employed.While convolutional neural networks (CNNs) have become the backbone of many image classification pipelines, they often struggle to capture global dependencies, especially in complex visual scenes like natural rock formations. TheConvNeXt architecture [1] has demonstrated state-of-the-art performance by combining the strength of hierarchical CNNs with training techniques inspired by Vision Transformers. However, its purely convolutional structure can limit long-range feature modelling. Vision Transformer models such as DeiT [2] excel at capturing global information but tend to underperform on small datasets and fine-grained tasks due to their data-hungry nature. Meanwhile, lightweight models like MobileNetV2 [3] are efficient but often trade off accuracy for speed and compactness.To overcome these limitations, we propose a modified ConvNeXt model enhanced with Self-Attention (SA) and Channel Attention (CA) modules. The self-attention mechanism captures non-local spatial dependencies, while channel attention adaptively reweights feature maps based on their relevance. This fusion

allows our model to better understand both the global context and fine-scale variations in rock images.

We benchmark our CNSCA model against ConvNeXt, MobileNetV2, and DeiT on a curated rock size classification dataset. The proposed model achieves an accuracy of 89.2%, outperforming ConvNeXt (82.1%), MobileNetV2 (64.%), and DeiT (82.2%). These results demonstrate the effectiveness of attention mechanisms in enhancing feature learning for fine-grained natural classification tasks.

2. DESCRIPTION OF MODELS

2.1. ConvNeXt: A Modernized Convolutional Neural Network

ConvNeXt is a convolutional architecture proposed by Facebook AI Research that revisits and upgrades traditional CNNs (like Residual Network-ResNet)[11] using design insights from Vision Transformers (ViTs). While it retains a fully convolutional structure, ConvNeXt improves performance to match or exceed that of transformer-based models—all without using self-attention. Below a summary of the key architectural innovations:

(i) Depthwise Separable Convolutions

ConvNeXt adopts depthwise separable convolutions, a technique popularized by MobileNet and later embraced in ViTs (via token mixing). Depthwise convolution applies a single convolutional filter per input channel. Pointwise convolution (1×1) then mixes information. This reduces computation while preserving performance. This design is structurally similar to MLP mixing in transformers.
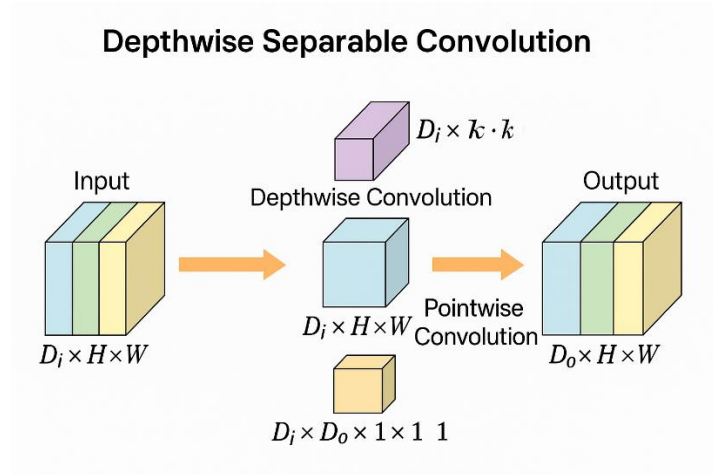

Figure 1 Depthwise Seperable Convolution

The cost of computation is $H \cdot W \cdot (D_i \cdot k^2 + D_i \cdot D_o)$ as compared to standard convolution $H \cdot W \cdot D_i \cdot D_o \cdot k^2$

(ii) Layer Normalization

Traditional CNNs use Batch Normalization, which depends on batch statistics. ConvNeXt replaces this with Layer Normalization, a choice inspired by transformer models. This is because

Layer Norm is more stable and compatible with non-sequential data and works better with large-scale training.

(iii) Large Kernel Sizes

Instead of small 3×3 kernels (common in ResNet), ConvNeXt uses 7×7 depthwise convolutions. This expands the receptive field and allows the network to capture broader spatial context, partially mimicking the global perspective of self-attention. This helps approximate the long-range interactions that transformers handle with self-attention—but using convolutions.

(iv) Gaussian Error Linear Unit

ConvNeXt replaces ReLU with GELU (Gaussian Error Linear Unit), which is the activation function used in BERT and other transformer models.GELU offers smoother and more expressive nonlinear behavior, which improves convergence and accuracy.

(v) ResNet-Like Stage Design with Transformer-

Like Refinementsthe overall structure still follows the stage-wise hierarchy of ResNet (e.g., 4 stages, downsampling between them).But block-level changes (e.g., inverted bottlenecks, norm-first ordering) are borrowed from transformers.

(vi) Residual Connections

Allow the network to bypass layers by adding the input of a block directly to its output. This helps train deep networks by preventing vanishing gradients and enabling feature reuse.

(vii) Inverted Bottleneck Structure (like MobileNet/Vit)

ConvNext expands the feature dimension first, processes it, then reduces it back to the original size.This allows richer feature transformation with low computational cost, while preserving important information (Figure 2).
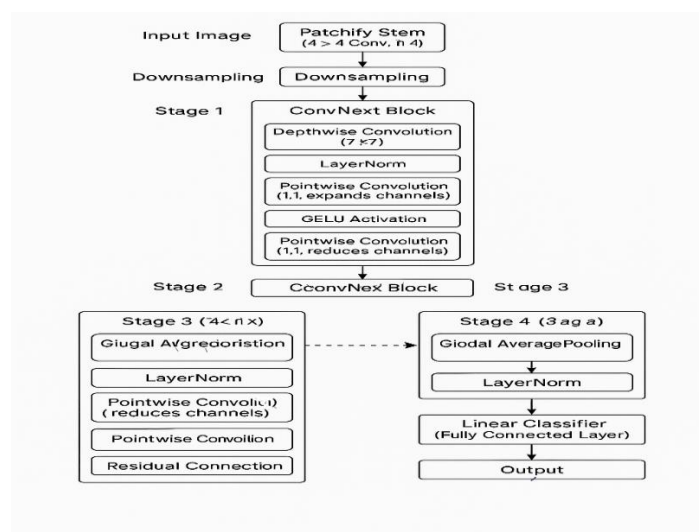

Figure 2ConvNext Architecture

2.2. ConvNext with Self and Channel Attention

Despite its many transformer-inspired components, ConvNeXtdoes not use self-attention or channel attention.
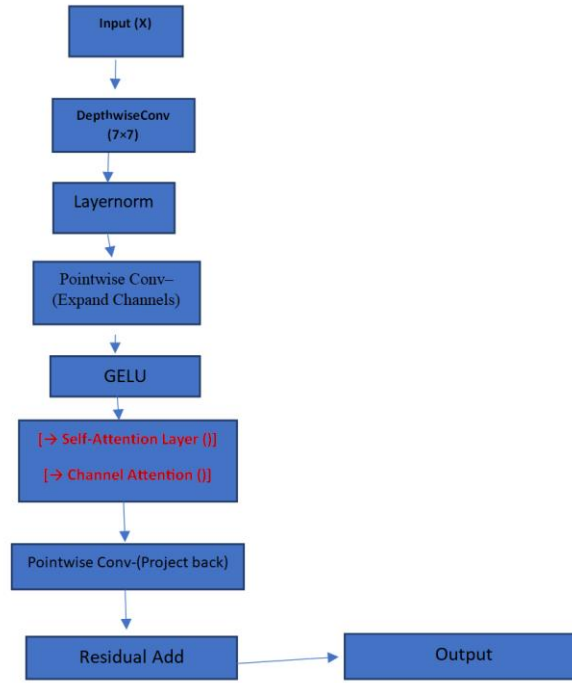

Figure 3 Proposed Block Model

This makes it faster and more computationally efficient but limited in modelling explicit global interactions, which transformers handle naturally. We propose adding self-attention and channel attention to ConvNext as shown in Figure 3.

Self-attention [4],[5] is a mechanism that allows each spatial position in a feature map to interact with all other positions, learning the relationships between distant regions in an image. This is especially useful for capturing global context and long-range dependencies that traditional convolutions struggle to model. It works by computing similarity scores (attention weights) between all positions and using them to weigh features accordingly. This improves the model's ability to reason about object structure and spatial layout.

Channel attention mechanisms [6], such as Squeeze-and-Excitation (SE) blocks, adaptively recalibrate the importance of feature channels by learning which ones contribute most to the task. This is typically done by applying global average pooling to compress spatial information, followed by a small neural network that generates channel-wise attention weights. These weights are then used to emphasize or suppressdifferent channels.Thus enhancing feature representation by focusing on the most informative filters.

3. EXPERIMENT AND RESULTS

To construct the dataset, a batch of industrial coal was manually sieved into two components: fines (<6 mm) and coarse particles (>6 mm). These components were then recombined in

controlled proportions to create seven distinct classes with fines content of 0%, 20%, 40%, 50%, 60%, 80%, and 100% as shown in figure 4. Each class was carefully prepared by blending the two components in the appropriate weight ratio `

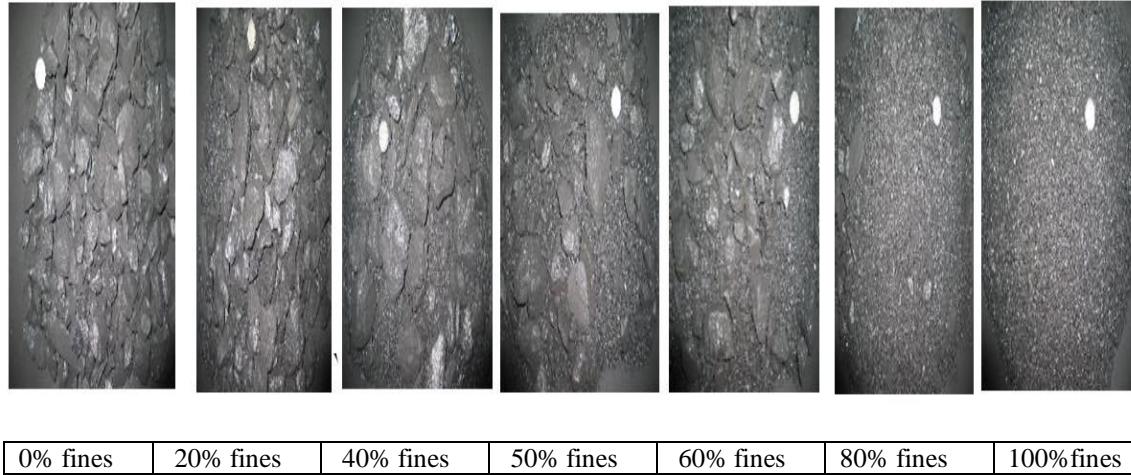| 0% fines | 20% fines | 40% fines | 50% fines | 60% fines | 80% fines | 100%fines |

Figure4. Examples of Dataset used for the experiment. Images of coal particle blends.

The percentage fines are indicated on bottom of each image. A 25 mm diameter South African five randcoin in the bottom three images gives a sense of scale.

To simulate industrial conditions, each prepared mixture was distributed onto **a** pilot-scale moving conveyor belt **e**quipped with a hopper. For each class, 10 high-resolution image**s** (2272×1704 pixels) were taken. To increase the dataset size and promote spatial diversity, each image was further split into four smaller sub-images (1132×832 pixels).Thus each class contains 40 images. Each image contains a South African R5 coin (diameter: 26 mm) as a visual reference for scale and particle size calibration (Figure 4).

To enhance the model's generalization capability and reduce overfitting, a series of data augmentation techniques were applied during training. These augmentations simulate natural variations in the dataset, such as differences in camera angle, lighting, and sample distribution. The augmentation pipeline included the following transformations: randomhorizontal flipping, random rotation, random zooming, random contrast adjustment and random translation. These operations were applied in real-time during training to each input image, resulting in a more diverse and robust dataset without increasing its physical size. The model was compiled with Adam optimizer with a learning rate of 0.0001, suitable for efficient gradient descent with adaptive learning. Sparse categorical cross entropy was used, indicating that the class labels are encoded as integers rather than one-hot vectors. Accuracy was used to monitor model performance during training and validation. The training was implemented with4 images per batch.10% of the data was allocated for validation.A seed value of 42 was used to ensure reproducibility of the training-validation split.
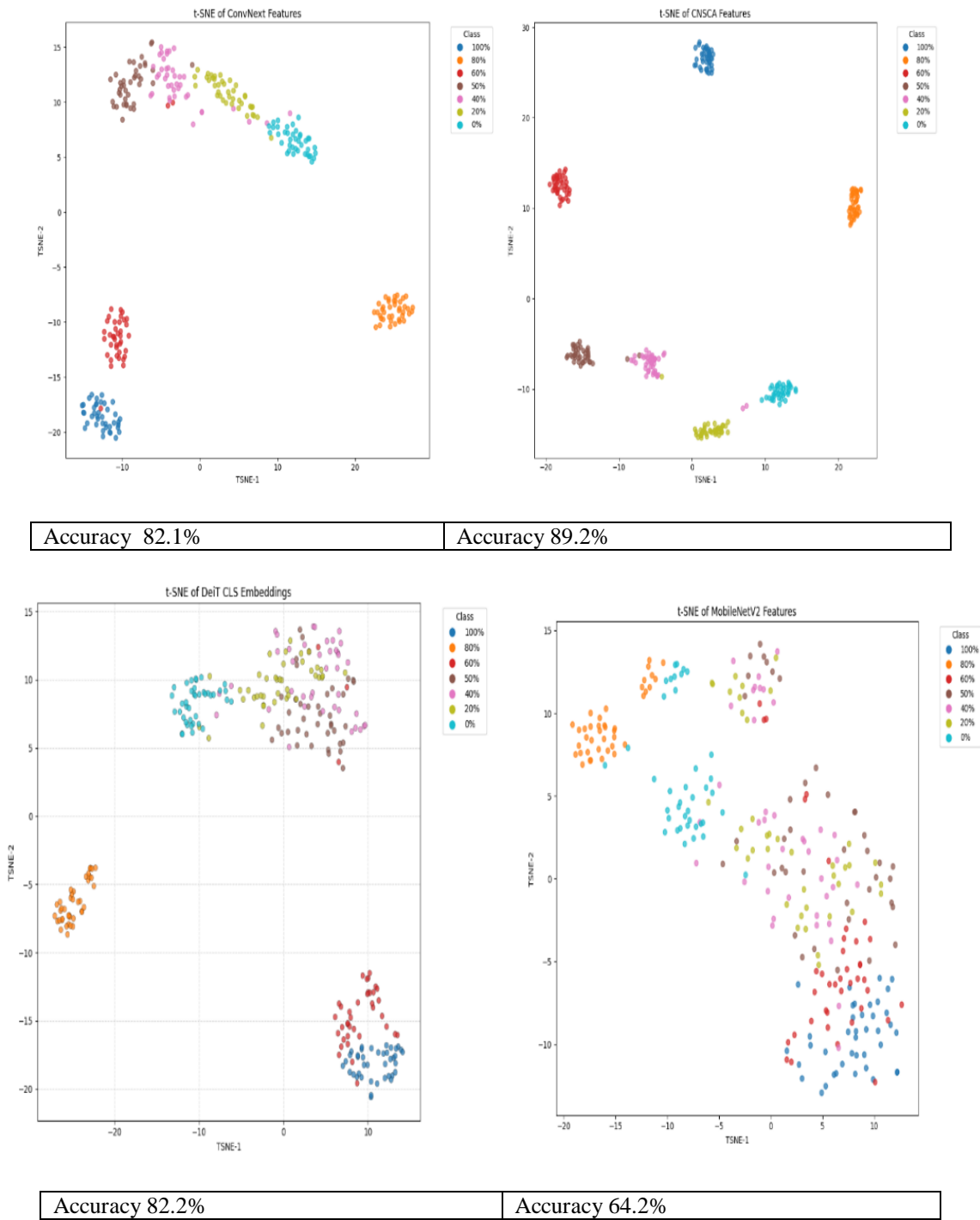
| Accuracy 82.1% | Accuracy 89.2% |

| Accuracy 82.2% | Accuracy 64.2% |

Figure 5 **t**-SNE score plots of feature representations of models. At the bottom are the corresponding accuracies.

Figure 5 presents the t-SNE score plots of feature representations extracted from various models, along with their corresponding classification accuracies. Each of the seven classes. The CNSCA shows the best separation of the seven classes. The least overlap between the 50% class and 40% class is seen in the CNSCA model. This is reflected in the corresponding accuracy which is 89.2%.

Figure 6 shows the confusion matrix which confirm the overlap between the 40% class and 50% class.
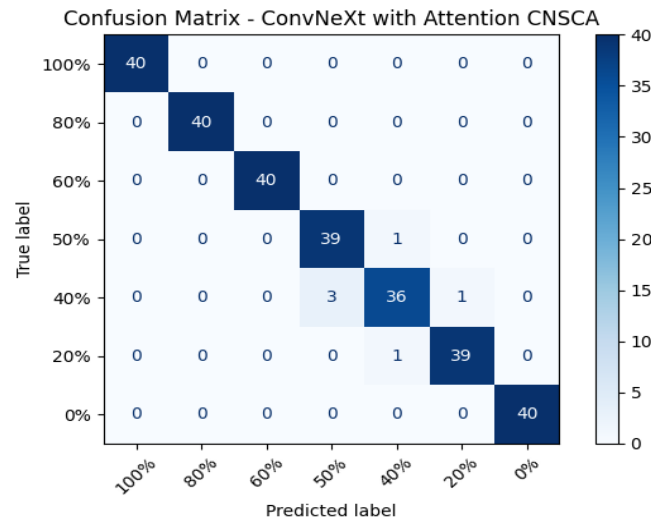

Figure 6 Confusion Matrix of proposed model

4. CONCLUSIONS

The findings highlight the effectiveness of integrating attention mechanisms into convolutional architectures for complex visual classification tasks. The proposed ConvNeXt-SA-CA model offers a promising approach for advancing automated rock size analysis, with potential applications across geotechnical engineering, mining operations, and resource management.

REFERENCES

[1]   Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11976–11986. https://doi.org/10.1109/CVPR52688.2022.01166

[2]   Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., &Jégou, H. (2021). *Training data-efficient image transformers & distillation through attention*. In Proceedings of the International Conference on Machine Learning (ICML), 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html

[3]   Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[4]   Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I*Attention is All You Need.*In Advances in Neural Information Processing Systems (NeurIPS), 30. . (2017). https://arxiv.org/abs/1706.03762

[5]   Wang, X., Girshick, R., Gupta, A., & He, K. *Non-local Neural Networks.*In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7794–7803.(2018). https://doi.org/10.1109/CVPR.2018.00813

[6]   Hu et al., 2018 – Introduced Squeeze-and-Excitation Networks (SE), which apply channel-wise attention.Hu, J., Shen, L., & Sun, G*Squeeze-and-Excitation Networks*. In Proceedings of the IEEE/CVF Conference on Computer Vision and PatternRecognition (CVPR), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745 . (2018).

[7]   A. Amankwah A C Aldrich "Estimation of particulate fines on conveyor belts by use of waveletand morphological image processing" Journal for Machine Learning and Computing   vol.1 no. 2 pp. 1-6 ISSN: 2010-3689 1 2011).

[8]   X Liu C Aldrich "Multivariate image processing in minerals engineering with vision transformers"- Minerals Engineering, 2024

[9]   A.Amankwah C. Aldrich"Automatic Rock Image Segmentation Using Mean Shift and Watershed Transform . Proceedings IEEE Radioelektonika 11, Brno Czech Republic April 2011

[10]   Lin, W., Li, X., Yang, Z., Lin, L., Xiong, S., Wang, Z., et al. (2018). A new improved threshold segmentation method for scanning images of reservoir rocks considering pore fractal

[11]   K He, Xi Zhang, S Ren, JSun "Deep Residual Learning for Image Recognition" Proceedings IEEE CVPR 2016

AUTHORS

Anthony Amankwah obtained his B.Sc. degree in Metallurgical Engineering from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, in 1996. He later earned B.Sc. and M.Sc. degrees in Electrical Engineering and Computer Science from the University of Duisburg-Essen, Duisburg, Germany, in 2003. He completed his Ph.D. in Electrical and Computer Science at the University of Siegen, Germany. He is currently employed in the Machine Vision industry in Germany

Chris Aldrich is a Fellow of the South African Academy of Engineering and currently serves as Professor and Chair of Process Systems Engineering at Curtin University, Australia. His research focuses on machine learning, data mining, and advanced process modeling and control systems, with applications in the chemical and mineral processing industries.