Waste Classification from Digital Images Using ConvNeXt

Jianchun Qi, Minh Nguyen, Wei Qi Yan

Auckland University of Technology, Auckland 1010 New Zealand

Abstract. In this paper, ConvNeXt is selected as a model for waste classification from digital images. ConvNeXt is a CNN-based backbone network that has been proposed to further improve the performance of models for visual tasks, following the various types of research work that have been generated based on Transformer. In this paper, we take ConvNeXt as the backbone to obtain an efficient waste classification model. In our experiments, we categorized waste into four classes based on predefined classification criteria. We have collected 1,660 labeled images for model training. By using ConvNeXt, we observed that the best experimental result in this paper was from ConvNeXt, which has achieved an accuracy 79.88% in the waste classification. In order to evaluate the model, we consider AP and mAP for waste classification. Our experimental results show that using Mask R-CNN network with ConvNeXt as the backbone outperforms the existing methods for waste classification.

Keywords: ConvNeXt · Mask R-CNN · Waste detection · Waste classification · Object detection

1 Introduction

An increasing amount of waste products are being consumed every day [3]. The disposal and recycling of wastes is a problem that should be considered in the process of protecting the environment. For waste classification, it should be grouped into categories according to its components, properties, value, and impact on the environment, depending on the type of disposal. In general, according to the characteristics of wastes, we group wastes into four major categories, namely hazardous waste, recyclable waste, wet waste, and dry waste. For example, after the waste is efficiently classified into the four classes and then transported to the waste treatment plant, the hazardous waste will be disposed to protect our environment. Recyclable waste will be sent to various resource recycling factories where it is recycled to save resources. Then, wet waste is tackled in the factory to produce biogas, which in turn helps to generate electricity and save electricity resources. The recycling of dry waste by incinerating it, can produce fuel for clean energy. If the waste is not classified and both dry waste and wet waste are treated at the same time, they are mixed and incinerated to produce significantly higher levels of carcinogenic substances, increasing the risk of secondary pollution to the environment [4]. We see that garbage

classification is of great significance for resource utilization and environmental protection.

At this stage, there are still many challenges in household waste classification, such as the lack of waste classification and the shortage of basic infrastructure for waste disposal. Therefore, an automatic waste classification method is of great value and significance to society. With the development of deep learning [11, 12, 14], it has become possible to improve the automated waste classification, realize the treatment and effective utilization of garbage. This provides a good chance for promoting the development of the municipal domestic waste treatment industry.

At present, a plethora of deep learning algorithms have achieved excellent results in waste classification [6, 28]. However, the characteristics of waste make the task of waste classification more difficult. For example, nut shells need to be classified into two classes: Dry garbage and wet garbage. Chestnuts need to be classified as wet waste, while walnut shells are dry waste. However, the similarity between the two types of nut shells is relatively high. During the experiment, we see that the two types of nut shells are very easily confused in the classifications.

Therefore, we aim to find a model that is much suitable for the task of garbage classification as a way to improve the accuracy of waste classification and save the classification costs. Furthermore, the collection of waste data is a major challenge due to the diversity of household waste, it is also a significant goal for us to collect waste images into a comprehensive, accurate, and diverse waste dataset. Overall, the main contributions of this paper are as follows:

(1) The training results of ConvNeXt model in waste classification have been conducted. The mAP of classification 79.88% was reached.

(2) Accurate waste images among a large number of waste images are found, a dataset with four classes of waste is constructed: Hazardous waste, dry waste, wet waste, and recyclable waste.

(3) The development of convolutional neural networks has a slew of advantages compared to attention mechanisms.

In this paper, we show our related work in Section 2, our methodology is depicted in Section 3, the result analysis is stated in section 4, and our conclusion will be drawn in Section 5.

2 Related Work

Convolutional Neural Networks (CNN or ConvNet) is a deep neural network with a convolutional architecture, which has the ability of representation learning [38]. It is usually composed of three parts. The first part is the input layer, the second is composed of multiple pooling layers and sampling layers, where the two types of layers, pooling and convolutional, are usually alternating, and the depth of each filter increases

sequentially from left to right. The final part then consists of one or more fully connected classifiers [39]. Among these, CNN has three critical operations, namely, local receptive fields, shared weights, and pooling layers, where CNNs have the advantage of reducing the number of network parameters and avoiding model overfitting [15, 16, 25].

Specifically, a convolutional layer of CNN contains a number of feature maps, each layer comprises a number of neurons. A neuron is only connected to the neurons in the adjacent layer and forms a rectangular arrangement of states [32, 37]. Neurons in the same layer then share the same weights. The convolutional kernel is initialized in the form of a random matrix, the weights are continuously trained during the training process for feature extraction. Besides, the pooling layer also has two forms: Average pooling and max pooling. This allows the complexity of the proposed model to be significantly simplified and the parameters of the model to be reduced.

Owing to the apparent advantages of convolutional neural networks, it has applications in face recognition, automatic speech recognition, gesture recognition, and natural language processing [30, 33]. Moreover, the performance of CNNs is even much outstanding for visual object recognition from digital images, as it allows images to be directly employed as the input to the network, and automatically extracts visual features, such as color and texture. Moreover, it avoids complex feature extraction and has excellent robustness and computational efficiency.

The depth of early CNN algorithms is critical to model performance, capable of extracting visual features. An increase in the number of network layers means that the network can extract more abundant features. Theoretically, the more layers a network has, the better the results will be. However, simply network leads to the vanishing and exploding gradient problem [22, 23, 37], which prevents the backpropagation process from effectively updating the gradients, resulting in the parameters not fully being updated. Currently, batch normalization can solve this problem [24, 31], which normalizes scattered data to prevent the gradient from vanishing in the process of backpropagation, so that the number of model layers can reach a level of dozens.

However, the increase in the number of network layers brings another problem, the degradation problem, in which the accuracy of network training saturates, and even performance degrade [9]. ResNet is also a relatively mainstream CNN algorithm at present. It introduces the residual network and makes use of a sliding window model to extract visual features and outputs a multilayer pyramidal feature map, which is ideal for a variety of downstream tasks. ResNet adds a shortcut [7] between every two layers of the network, forming a residual network. All residual blocks do not have a pooling layer, and directly use the convolution with stride 2 for downsampling. This residual network structure allows each layer of the neural network to fully learn the residuals of the previous layers' output and preserve the integrity of the information.

Besides ResNet, there are a consortium of CNN-based models, such as Mask R-CNN [8] and YOLO series [1, 40]. Mask R-CNN is a two-stage model. The first stage generates proposals, the second stage generates masks and bounding boxes. However, the backbone network of Mask R-CNN is ResNet-50, a Feature Pyramid Network (FPN) is introduced so that the feature map of each layer can be fused to extract the features of each layer. Moreover, Mask R-CNN also introduces a mask branch. As a result, Mask R-CNN has the advantage of both high speed and high accuracy. There are a series of YOLO models, the latest one is YOLOv7 [35]. YOLO is a one-stage model [19, 20, 26, 36], which do not show the step of generating proposals and directly take use of regression method for detection.

Transformer is the first deep learning model that is based entirely on attention mechanism to improve the speed of model training [34], which is primarily proffered in the field of natural language processing (NLP) [27]. It is structured by using encoder and decoder. Among them, self-attention layer and multihead attention mechanism are the key parts of this model. Therefore, its advantage is that it can be efficiently parallelized. After that, the attention mechanism was gradually developed, the attention-based models in the field of vision have been proposed, such as Vision Transformer [5], Swin Transformer [17], and DERT [2].

CNN has a slew of advantages for extracting visual features, such as detecting the boundary of visual objects and other basic visual elements. However, because attention-based models have an attention mechanism, they are able to capture global contextual information, which has a larger receptive field and substantial model representation capabilities. We see that attention-based models are much more effective in dealing with high-level visual effects.

Therefore, we speculate that the combination of the two advantages of CNN model and the attention-based model may have better development prospects. This is also one of the purposes of this paper to prob the newly proposed ConvNeXt network.

3 Our Method

In the last two years, attention-based models have become mainstream research, such as Swin Transformer. Although Swin Transformer has achieved remarkable achievements for various vision tasks and successfully solved the problem of substantial computational cost, there is still room for improvement. For example, its calculations based on sliding windows are much complex, which makes it challenging to be deployed. With recent findings, the depthwise convolution appears to be equivalent to the self-attention mechanism. In particular, the performance of depthwise convolution is deeper than that of Swin Transformer in the case of smaller parameters.

Hence, pertaining to waste classification, we concentrate our research intention on the newly proposed ConvNeXt net [18], which outperforms Swin Transformer to explore the most suitable algorithm for waste classification.

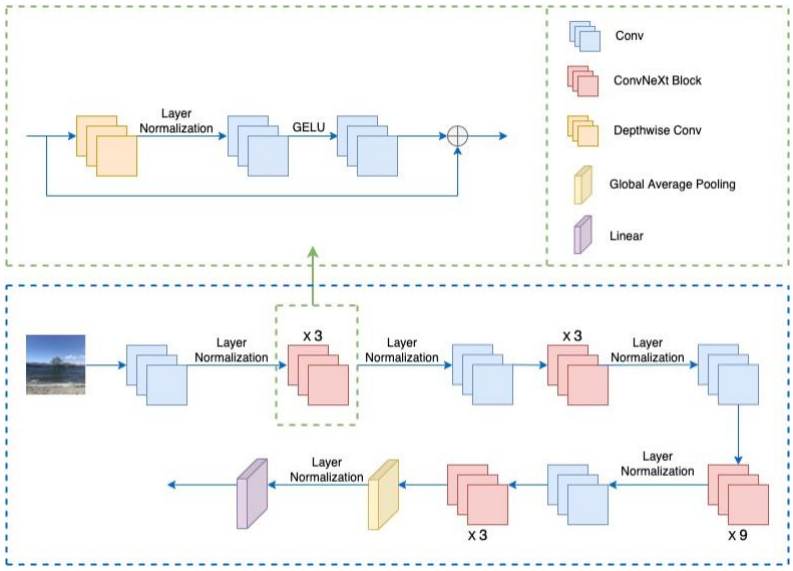

Fig. 1. The framework of our model.

ConvNeXt outperforms Swin Transformer in terms of not only accuracy but also performance and simplicity. ConvNeXt is based on ResNet structure and improves the model in five main ways:

(1) Macro design replaces the stem cell layer of ResNet with a Patchify layer and adjusts the computational ratio to approximately 1:1:3:1.

(2) In ResNeXt, depthwise convolution [4] is employed, the network width is set to 96.

(3) Related to inverted bottleneck, the inverted bottleneck in the Transformer block is adopted.

(4) With very large kernel size, a larger convolution kernel is taken.

(5) Under the assistance of various layer-wise micro designs, ReLU was replaced with GELU. Furthermore, fewer activation functions and fewer layer normalization were used.

The network structure of ConvNeXt is shown in Fig. 1. We combine ConvNeXt with Mask R-CNN. Mask R-CNN is chosen because it has an excellent performance in downstream tasks which is a very flexible framework that can add multiple branches to accomplish various tasks, such as instance segmentation, semantic segmentation, and human pose recognition. In ConvNeXt, stride is 4. The numbers of channels are 96,

192, 384, and 768, which correspond to the numbers of blocks stacked in each stage, i.e., 3, 3, 9, and 3. After that, layer normalization and GELU are added to the net. Finally, drop path is adopted to prevent overfitting and improve performance.

In Mask R-CNN, five loss functions are employed, which contain two loss functions for the RPN network [8], two loss functions for classification, and one loss function for the mask branch. The first four loss functions are as same as the loss function of Faster R-CNN [29], the final mask loss function adopts the mask branch with Km^2 output for each ROI. By using a per-pixel sigmoid, the average binary cross-entropy loss is obtained for the pixels. In this way, the model only needs to detect which class the ROI is, which only calculates for one branch and avoids the competition among classes.

4 Result Analysis

4.1 Our Dataset

There are various types of domestic wastes, which are classifiered into four main classes according to the waste classification criteria, namely, dry waste, wet waste, hazardous waste, and recyclable waste. Within each class, there is a consortium of sub-categories. For example, recyclable waste includes cardboard, glass, and plastic bottles. The hazardous waste includes batteries, nail polish bottles, and medicine bottles. Each of these classes needs to be collected in sufficient quantities. Therefore, the diversity of waste classes results in a challenging task to collect the waste dataset.

In our experiment, we collected a total of 1,660 images. For each class, the number of samples is around 400. Besides, during the training process, our dataset consists of training, validation, and test sets with sample sizes of 1,328, 166, and 166, respectively. Table 1 shows the classes of the waste dataset.

A group of representative samples of the datasets are shown in Fig. 2. While collecting waste data, we considered the diversity of the samples. In our dataset, the detected objects have various shapes, Fig. 2(g) shows a plastic bottle. In addition, there are also plastic bottles presented in the dataset from various angles. Other samples were selected in the same way. We will also select images that show only a portion of the object as our visual data.

Table 1. The number of waste dataset

Classes	Numbers
Dry waste	460
Wet waste	380
Hazardous waste	450
Recyclable waste	370
TOTAL	1,660

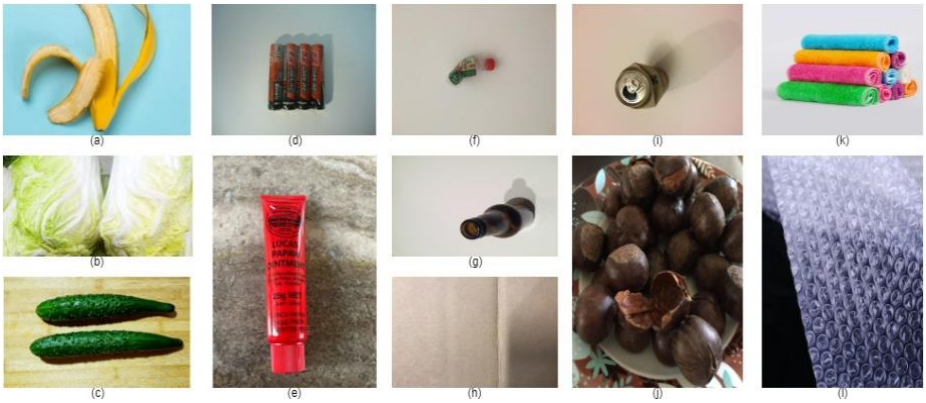

Fig. 2. The samples of the waste dataset. (a), (b), (c), and (j) are banana peel, cabbage, cucumber, and chestnut shell, respectively, which belong to wet class. (d) and (e) are battery and ointment shell, respectively, which are classified to hazardous class. (f), (g), (h), and (i) are a plastic bottle, glass bottle, cardboard, and can, respectively, which are grouped to recyclable class. (k) and (l) are rag and bubble film, respectively, are identified as dry class.

After the data had been collected, we annotated each sample by using the labeling software LabelMe. Our annotations follow the JSON file format of the COCO dataset. Then, we labelled the four types of images as “Hazardous”, “Dry”, “Wet”, and “Recyclable”, respectively. Finally, we save the images in JPG format.

4.2 Evaluation Methods

In order to better improve the model, we need to evaluate its performance of the model. According to our experimental purpose, we choose to use average precision (AP) and mean average precision (mAP).

Accuracy is a popular evaluation metric, which is a number of correctly selected samples divided by using the number of all samples. General speaking, the higher the accuracy rate, the better the classifier. Eq. (1) shows the algorithm for accuracy.

$$\textbf{Accuracy} = (TP+TN)/(TP+FP+TN+FN) \tag{1}$$

where *TP*, *TN*, *FP*, and *FN* are true positive, true negative, false positive, and false negative. Positive and negative represent the predicted results, positive samples are predicted to be positive, and negative samples are predicted to be negative. True or false indicates the ground truth of samples, positive or negative refers to the test results.

TP+TN is the sum of all correctly predicted positive samples and negative samples, *TP+TN+FP+FN* is the total number of samples.

Besides, precision measures the probability of a classifier judging a positive sample to be a true positive sample. AP is the area under the curve, and mAP refers to the average of multiple classes of precisions. The calculation method of precision is shown in Eq. (2).

$$\textbf{Precision} = TP/(TP+FP) \tag{2}$$

4.3 Result Analysis

Fig. 3 shows us an example of our waste detection results. We see that there are various classes of waste samples in the image, each class has the colored bounding box and label.

In Fig. 4, we see that AP rates of the four classes: Dry, wet, recyclable, and hazardous, are 59.90%, 81.51%, 95.50%, and 82.61%, respectively. Hence, Fig. 5 shows the mAP rate is 79.88%.

We quantitatively compare our model with other models by using our own dataset. Table 2 shows the comparison results. The accuracy of Swin Transformer is 0.78, which is slightly lower than that of the ConvNeXt by 0.02. The accuracy rates of YOLOv3 and ResNet-50 are 0.77 and 0.75, respectively. It is only 0.05 different from ConvNeXt most. This means that the ConvNeXt model is currently the better performing model, but could be further improved.

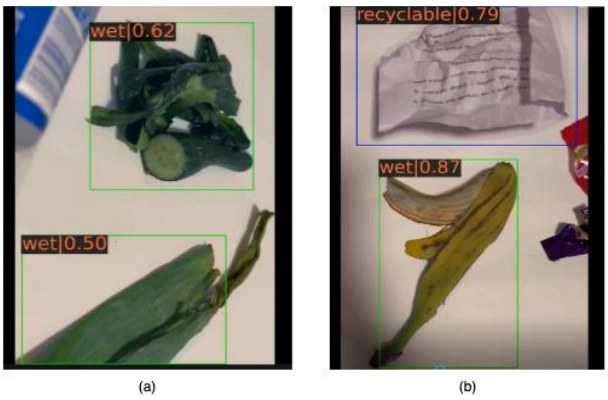

Fig. 3. ConvNeXt-based classification results from digital videos. (a) The results of classifying green onion leaves and cucumber peel are classified to the class wet. (b) The classification results of banana peel and waste paper, which are classified to class wet and class recyclable, respectively.

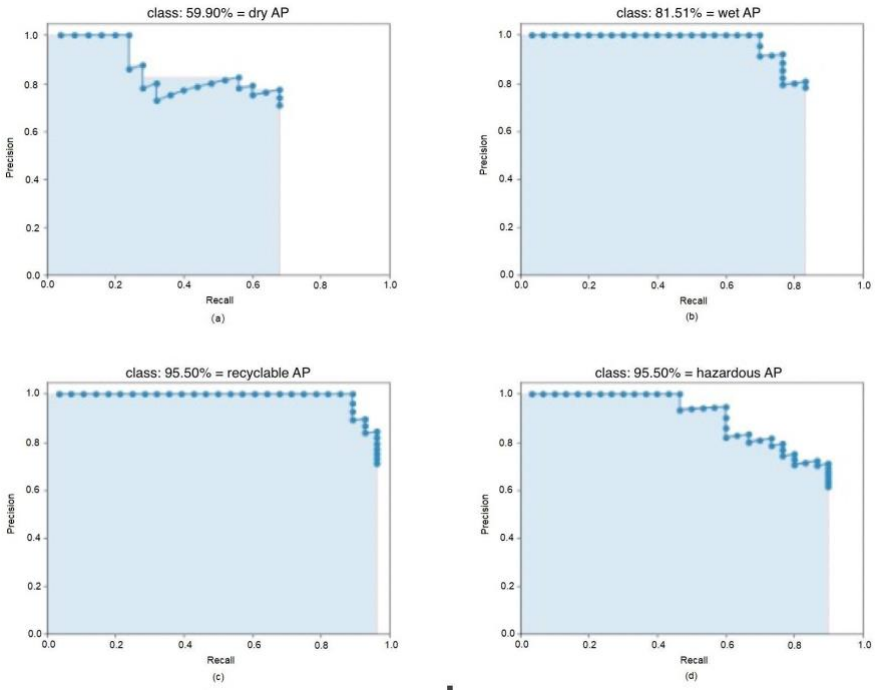

Fig. 4. Average precisions of the four classes classification. (a) The average precision of class dry. (b) The average precision of class wet. (c) The average precision of class recyclable. (d) The average precision of class hazardous.

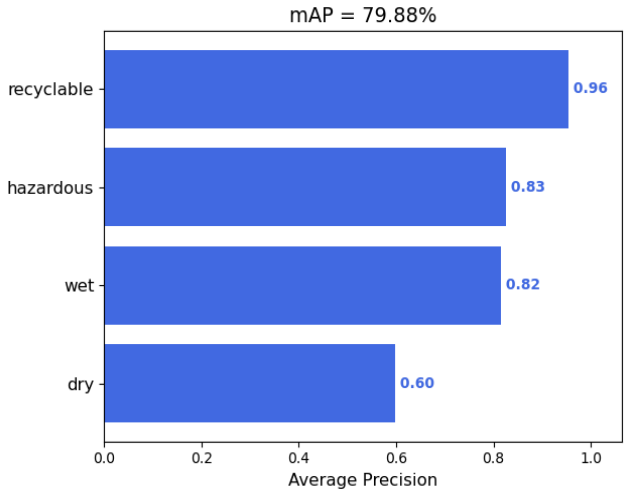

Fig. 5. Mean average precisions of the four classes.

Table 2. Mean average precision results between four models

Dataset	ConvNeXt	Swin Transformer	YOLOv3	ResNet-50
Recyclable waste	0.96	0.91	0.81	0.83
Hazardous waste	0.83	0.80	0.72	0.79
Dry waste	0.60	0.65	0.67	0.58
Wet waste	0.82	0.76	0.88	0.80
AVERAGE	**0.80**	**0.78**	**0.77**	**0.75**

Table 3. Mean average precision results between the three algorithms

Model	mAPs
ConvNeXt+ Mask R-CNN	0.80
ConvNeXt+ Cascade Mask R-CNN	0.79
ConvNeXt+ Faster R-CNN	0.76

Table 4. Mean average precision results of the three backbones

Model	mAPs
ConvNeXt+ Mask R-CNN	0.80
Swin Transformer+ Mask R-CNN	0.78
ResNet-50+ Mask R-CNN	0.73

Furthermore, we kept the ConvNeXt as the backbone network approach to whether Mask R-CNN was the best performing model when combined with ConvNeXt. As shown in Table 3, by using Cascade Mask R-CNN in combination with ConvNeXt, the mAP dropped 0.01. After replacing Mask R-CNN with Faster R-CNN, the mAP was the lowest one, up to 0.76.

Finally, we experimented with other backbone networks instead of ConvNeXt and obtained Table 4. After substituting the backbone network with Swin Transformer and ResNet-50, the mAP values dropped by 0.02, and 0.07 to 0.78 and 0.73, respectively. With this experiment, we see that the ConvNeXt model has the most stable performance.

4.4 Ablation Experiments

To gain a deeper understanding of the ConvNeXt model, we carried out a number of ablation experiments. Pertaining to ConvNeXt, we conduct comprehensive ablation studies on the model through three aspects.

4.4.1 Activation Functions

One of the indispensable features in the training process is nonlinearity. At the same time, for the generalization ability of the model, random regularization, such as dropout,

needs to be added. Activation functions play an essential role for the proposed models. The nonlinear properties are introduced into the network which solves problems that cannot be solved by linear models. GELU [10] introduced the idea of random regularization and outperformed ReLU experimentally. Therefore, in the ConvNeXt net, the ReLU activation function was replaced by the GELU activation function in the ConvNeXt Block. However, ConvNeXt harnesses only one GELU activation function in each block.

Hence, in our experiments, we added a GELU to the convolutional layer in ConvNeXt Block as a way to investigate the effect of the number of GELU activation functions on the model. In Table 5, we see that after changed the number of GELU to two, the AP value of the model drops by 0.50 to 0.68, which shows that not the more activation functions are, the better the model performance is.

Table 5. Influence of activation function on average precision values

GELU	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
× 1	0.73	0.74	0.69	0.28	0.56	0.76
× 2	0.68	0.71	0.62	0.25	0.50	0.72

4.4.2 Batch Normalization

In convolutional neural networks, normalization has an important significance in preventing gradient disappearance and gradient explosion. At present, Batch Normalization (BN) is widely employed, while ConvNeXt makes use of Layer Normalization (LN). Hence, we investigate the impact of batch normalization on the performance of the model. As shown in Table 6, we see that replacing LN with BN is an imperfect solution, which results in a slight decrease of 0.02 in AP value from 0.73 to 0.71.

Table 6. Impact of Batch Normalization on average precision rates

LN	BN	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	√	0.71	0.73	0.65	0.26	0.52	0.75
√	×	0.73	0.74	0.69	0.28	0.56	0.76

4.4.3 Layer Normalization

Previously, we replaced layer normalization with batch normalization and received the result that the value of AP dropped. We investigated the number of LN again. In ConvNeXt, only one layer normalization is employed in the block. In Table 7, we have experimented with two-layer normalizations and three-layer normalizations, respectively, the results show that the more layer normalizations are, the smaller the AP value is. However, the AP value decreases slightly; it seems that the number of layer normalizations does not significantly affect the model performance.

Table 7. Influence of Layer Normalization on average precision rates

LN	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
× 1	0.73	0.74	0.69	0.28	0.56	0.76
× 2	0.71	0.73	0.67	0.26	0.55	0.76
× 3	0.71	0.72	0.64	0.25	0.54	0.75

Based on these three ablation experiments, we see that the number of activation functions has significant impact on the performance of ConvNeXt. The AP rates differed substantially from the increase of the GELU. Conversely, the effect of changing the number of LNs on the model performance is insignificant. These indicate that the activation function plays a decisive role in the performance of the ConvNeXt.

The experimental results show that the algorithm has stable performance, the design of model structure is much simpler than that of Swin Transformer. ConvNeXt combines the excellent components of high-performance neural network models, improves the performance of CNN to 87.8%, and creates inspiration for our research outcomes based on the attention-based models.

5 Conclusion

In this paper, we propose to utilize a network based on ConvNeXt as the backbone network for waste classification. It was constructed to conduct an efficient waste classification by using deep learning method. Overall, 1,660 images were collected as the waste dataset. We manually annotated these images. The experimental results show that the accuracy of ConvNeXt is 79.88%, which has a better performance compared to other peer models.

We see that in the process of the rapid development of attention mechanism, there is also a room for the exploration of convolutional neural network. Further research work on convolutional neural networks is needed. It is also indispensable to improve the accuracy and efficiency of waste classification. Finally, the collected waste dataset needs to be updated.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, M.Y.: YOLOv4: Optimal speed and accuracy of object detection. arXiv (2020).
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. ECCV, pp. 213-229 (2020).
3. Chen, S.S., Huang, J.L., Xiao, T.T., Gao, J., Bai, J.F., Luo, W., Dong, B.: Carbon emissions under different domestic waste treatment modes induced by garbage classification: Case study

in pilot communities in Shanghai, China. Science of the Total Environment. 717, 137193 (2020).
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. IEEE Conference on Computer Vision and Pattern Recognition (2017).
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X.H., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv (2020).
6. Funch, O.L., Marhaug, R., Kohtala, S., Steinert, M.: Detecting glass and metal in consumer trash bags during waste collection using convolutional neural networks. Waste Management. 119, 30-38 (2021).
7. Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence 2. 11, 665-673 (2020).
8. He, K.M., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. IEEE ICCV, pp. 2961- 2969 (2017).
9. He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J.: Deep residual learning for image recognition. IEEE CVPR, pp. 770-778 (2016).
10. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv (2016).
11. Ji, H., Liu, Z., Yan, W., Klette, R.: Early diagnosis of Alzheimer's disease based on selective kernel network with spatial attention. Asian Conference on Pattern Recognition 2 (1), 503-515 (2019).
12. Ji, H., Liu, Z., Yan, W., Klette, R.: Early diagnosis of Alzheimer's disease using deep learning. ACM ICCCV (2019).
13. Kang, Z., Yang, J., Li, G.L., Zhang, Z.Y.: An automatic garbage classification system based on deep learning. IEEE Access. 8, 140019-140029 (2020).
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM. 60, 84-90 (2017).
15. Liang, S. Yan, W.: A hybrid CTC+Attention model based on end-to-end framework for multilingual speech recognition. Multimedia Tools and Applications (2022).
16. Liu, X., Neuyen, M., Yan, W.Q.: Vehicle-related scene understanding using deep learning. ACPR 2019. CCIS, vol. 1180, pp. 61-73. Springer (2020).
17. Liu, Z., Lin, Y.T., Cao, Y., Hu, H., Wei, Y.X., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. IEEE ICCV, pp. 10012–10022 (2021).
18. Liu, Z., Mao, H.Z., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie. S.N.: A ConvNet for the 2020s. arXiv, (2022).
19. Luo, Z., Nguyen, M., Yan, W.: Kayak and sailboat detection based on the improved YOLO with Transformer. ACM ICCCV (2022).
20. Luo, Z., Nguyen, M., Yan,W.: Sailboat detection based on automated search attention mechanism and deep learning models. IEEE IVCNZ (2021).
21. Nie, Z.F., Duan, W.J., Li, X.D.: Domestic garbage recognition and detection based on Faster R-CNN. Journal of Physics: Conference Series (2021).
22. Nixon, M., Aguado, A.: Feature Extraction and Image Processing for Computer Vision. Academic Press (2019).

23. Pan, C., Yan, W.Q.: Object detection based on saturation of visual perception. Multimed. Tools Appl. 79(27-28), 19925-19944 (2020).
24. Pan, C., Liu, J., Yan, W., et al.: Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing, 30, 4773 - 4787 (2022)
25. Pan, C., Yan, W.: A learning-based positive feedback in salient object detection. IVCNZ, pp. 311-317 (2018)
26. Prince, S.J.: Computer Vision: Models, Learning, and Inference. Cambridge University Press, Cambridge (2012).
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI, 1. 8, 9 (2019).
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. IEEE CVPR, pp. 779-788 (2016).
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS 28 (2015).
30. Sakalle, A., Tomar, P., Bhardwaj, H., Acharya, D., Bhardwaj, A.: A LSTM based deep learning network for recognizing emotions using wireless brainwave driven system. Expert Systems with Applications. 173, 114516 (2021).
31. Shen, D., Xin, C., Nguyen, M., Yan, W.: Flame detection using deep learning. International Conference on Control, Automation and Robotics (2018).
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv (2014).
33. Srivastava, N., Geoffrey, H., Alex, K., Ilya, S., Ruslan S.: Dropout: a simple way to prevent neural networks from overfitting. J. Mac. Lear. 15, 1929-1958 (2014).
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2019).
35. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv (2022).
36. Xiao, B.J., Minh N., Yan, W.Q.: Apple ripeness identification using deep learning. International Symposium on Geometry and Vision. 1386, 53 (2021).
37. Xin, C., Nguyen, M., Yan, W.: Multiple flames recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 296-307 (2020).
38. Yan, W.Q.: Computational Methods for Deep Learning - Theoretic. Practice and Applications. Springer, Heidelberg (2021).
39. Yan, W.Q.: Introduction to Intelligent Surveillance - Surveillance Data Capture, Transmission, and Analytics, 3rd edn. Springer, Heidelberg (2019).