Oil Palm Fruit Image Ripeness Classification with Computer Vision using Deep Learning and Visual Attention

Herman[1,3], Albert Susanto[1,3], Tjeng Wawan Cenggoro[2,3], Suharjito[1] and Bens Pardamean[1,3]
[1]*Computer Science Department, BINUS Graduate Program – Master in Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480.*
[2]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480.*
[3]*Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia 11480*.
herman005@binus.ac.id

Abstract— Oil palm is one of the leading agricultural industries, especially in the South East Asian region. However, oil palm fruit ripeness classification based on computer vision has not gained many satisfactory results. Therefore, most of the ripeness sorting processes are still done manually by labor works. The objective of this research is to develop a model using a residual-based attention mechanism that could recognize the small detail differences between images. Thus, the model could classify oil palm fruit ripeness better. The dataset consists of 400 images with seven levels of ripeness. Since the number of images in the dataset, Ten Crop preprocessing is utilized to augment the data. The experiment showed that the proposed model ResAtt DenseNet model, which uses residual visual attention could improve the F1 Score by 1.1% compared to the highest F1 Score from other models in the experiment of this study.

***Index Terms*— Computer Vision; Convolutional Neural Network; Oil Palm Fruit Classification; Visual Attention.**

I. INTRODUCTION

Agriculture is one of the most vital industries in Indonesia. Therefore, numerous research have been conducted to support the productivity of the agriculture industry [1], [2]. One of the largest agriculture industries in Indonesia is an oil palm plantation. The amount of exported oil palm fruit has been increasing each year [3]. Unfortunately, most of the palm fruit sorting processes are still done manually because of the limited reliable system for oil palm fruit classification. Oil palm fruit has a complex appearance; hence, it is difficult to classify the level of ripeness as it needs to distinguish from the fruit in a group of thorns. A more ripe oil palm fruit has lesser thorns, and the pattern of how the colors spread is used to determine how ripe the fruit is. Numerous research have been conducted for oil palm fruit classification tasks, using machine learning or deep learning [4]–[9]. However, most of the previous proposed methods have not yet explored the advantages of deep learning methods. Instead, it relies heavily on manual processes. One of the most robust approaches for automatic fruit grading is by using a deep-learning-based model for oil palm fruit image classification. Deep learning has shown a promising result for image classification task even for a complex case [10]. Therefore, it is a promising candidate for the case of oil palm fruit classification. The utilization of a standard deep learning model for oil palm fruit classification has been studied by Ibrahim et al. [11].

The specific type of deep learning model, the Convolutional Neural Network (CNN), was first introduced in 1987 [12], but since the technologies at that time were limited, it only started to be popular in 2012 after AlexNet [10], a CNN with deeper layer than the original counterpart, won the ILSVRC-2012 image classification competition. Since the resurgence, numerous studies showed improved performance on the ImageNet dataset [13], [14] as the models get more layers. For instance, the ResNet model that can yield up to 151 layers [15]. Following the trend, some of the latest state-of-the-art models are based on the Densely Connected Network (DenseNet) [16], in which the smallest variation comprises 121 layers.

Unfortunately, the available oil palm fruit image datasets are relatively small, which is unfavorable for the standard deep learning models. It has been known that the standard deep learning models can only achieve its optimal performance with a large dataset. Therefore, a more complex and deeper architecture is necessary to be developed. To gain a robust performance on a small dataset, the complex model can be tailored to extract unique features from the oil palm fruit image such as how the ripeness colors spread or what is the pattern of the spikes. For the model to be able to focus on the unique features, one of the possible approaches is to use the attention mechanism.

The earliest study of attention mechanism was done in 2011 by Evans [17], in which he explains that human tendency to focus on some important part of their visual projection while blurring the less important part, the goal of implementing visual attention is to allow the machine to see which part of the input object is more relevant, when it is used for training during the classification or task specified. An illustration of the attention mechanism in image classification can be seen in Figure 1. The image in Figure 1 is taken from https://farm3.static.flickr.com/2337/2387965665_1d4278c661.jpg. Follow-up studies showed that implementing the attention mechanism on CNN can help the model to perform better on Computer Vision tasks [18]–[20]. Therefore, we proposed the use of the attention mechanism for the oil palm images classification problem. The proposed attention mechanism is a novel module that is based on the residual connection [15]. The experiments showed that the novel attention mechanism performed better than the standard attention mechanism that is typically implemented by using Sigmoid activation function. We also compared the proposed

models with the AlexNet model used by Ibrahim et al. [11] for oil palm image classification and the DenseNet with Squeeze and Excitation (SE) layer that can be interpreted as an attention mechanism. It is currently the state-of-the-art model for the Chest X-Ray image classification problem.

Figure 1: Eggplant detection attention

II. Related Works

The attention mechanism has been used in several neural network models to increase the performance and to help in specifying the important part of the image that is relevant to the task. The first attempt to apply the attention mechanism into deep learning was to generate a caption from the image [18] based on the Recurrent Neural Network. It introduced two kinds of attention: A soft attention mechanism, which attends to multiple objects in the image; and a hard attention, that focuses only on one area. Since then, a multiple attention based research follows, such as a network for speech recognition [21] that calculates the Euclidean loss of attention product from convolution, and a saliency map for joint learning during training, detecting multiple object on different scales [22]. Using an image saliency mapping detector, the saliency map was done by creating a feature map using a convolution method before feeding it forward to RNN repeatedly to generate a caption based on different attention part every time, and cropping important part of the picture for plant classification so that the network only see the relevant things [23].

The multiple methods that extend the model architecture, which is similar to the attention-like mechanism are Feature Pyramid Network [24]. This network scales all of the previous convoluted layer and feeds it forward with a convoluted tensor from the previous layer, making it looks like top-down pyramid . It then scales back from those result for object detection. The Squeeze-and-Excitation Networks (SENet) also locate areas of the original feature map by squeezing the global feature into one layer, and then scale it back to their original sizes [25]. While most of proposed methods for oil palm fruit image classification doesn't add any modification on the neural network, the networks have a harder time to classify an image that only have differences involving colors, spike, or any other small detail features. Thus, the models can be improved further.

The models proposed in this paper are also inspired by the concept of Skip-net [26], which uses feature concatenated from multiple fully connected network results on each convolution layer in the middle of the network, Share-net that read multiple sizes of the same input under the same weight across the neural networks [27], and Residual Network [15] by connecting all information on the previous layer residually.

A. AlexNet

AlexNet was first introduced in 2012 by Alex Krizhevsky [28], and it was considered to be the first convolutional neural network that was able to classify the object using the concept of neural networks and able to outperform other proposed methods during the competition. Using the computation power from Graphical Processing Unit (GPU), the potential of the convolutional neural network could finally be used when using a Computer Processing Unit (CPU) is not enough or taking a very long time to train a single model.

Table 1 shows all the layers of AlexNet since the first architecture for Convolutional Neural Network (CNN). Nowadays, the architecture is not considered deep, and it is mostly outperformed by other new architecture that uses residual, inception, or densely connected neural networks. However, the concept still remains as it uses a combination of convolution, activation, and fully connected layer. Since the architecture is simple, it is popular for neural network newcomers to tweak around the CNN model for classification tasks.

Table 1
AlexNet Architecture

Layer Name	Size
Conv2D ReLU	11×11
MaxPool2D	3×3
Conv2D ReLU	5×5
Conv2D ReLU	3×3
Conv2D ReLU	3×3
Conv2D ReLU	3×3
MaxPool2D	3×3
AvgPool2D	6×6
Fully Connected ReLU	4096
Fully Connected ReLU	4096
Fully Connected	{num_of_classes_to_predict}

B. DenseNet with SE

Squeeze and Excitation (SE) layer is originally an extension proposed for ResNet that uses a block to squeeze the global feature of the previous layer into a single feature and multiplies it with the original layers so that the model can decide the importance from features of each channel. The ResNet + SE Layer is called as SENet in the original paper [25]. The main appeal of this method is the simple implementation that works on almost all types of convolutional modules [25]. For instance, it can be implemented on the DenseNet block, as seen in Figure 2.

Although it is not specifically proposed as an attention mechanism, SE Layer can be viewed as a layer that provides attention. The attention generated from SE Layer is an element-wise-multiplication with the output of the convolutional module it is attached to.

The SENet has been improved further by Yan et al. [29] for the Chest X-Ray image classification problem. They found that the use of DenseNet as the SENet backbone can achieve the state-of-the-art performance in Chest X-Ray 14 dataset [30]. This finding is consistent with the fact that DenseNet performs well in transfer learning for the Chest X-Ray images problem [31]. The full architecture of DenseNet + SE Layer is illustrated in Figure 3.

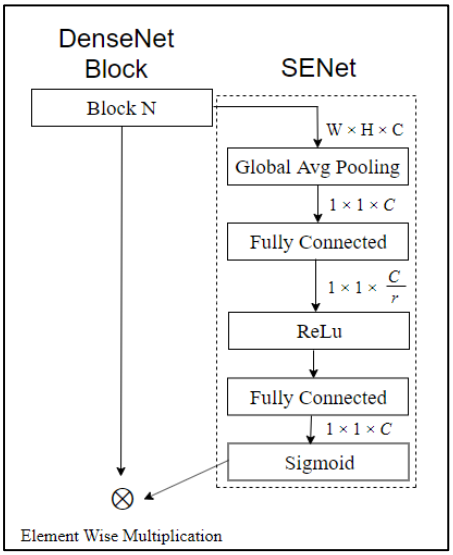


Figure 2: DenseNet Block + SE layer [25]

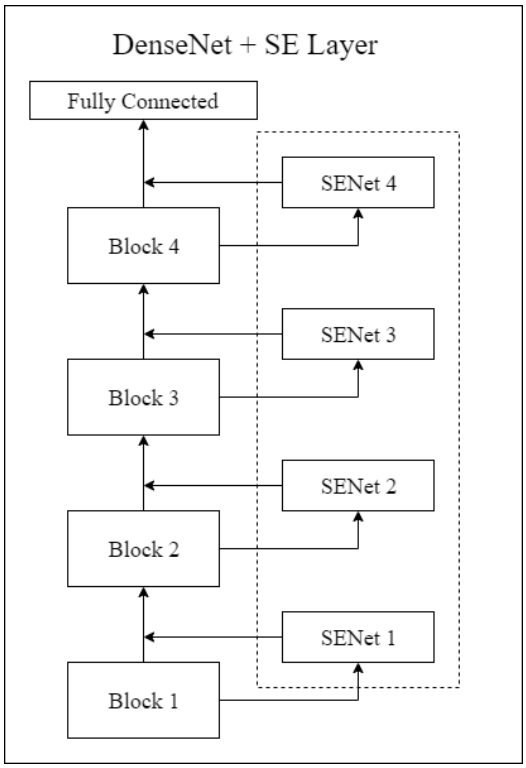


Figure 3: DenseNet + SE layer architecture

III. Residual Attention

The purpose of implementing an attention module on every DenseNet block is to let the network to perceive better on the different parts during feature extraction. By implementing a simple convolution at the end of each block, it could extract the important parts of the feature maps. However, this design could lead to a vanishing gradient problem, as the typical attention module used Sigmoid activation function. Since the general idea of visual attention is to focus on an important feature and multiply it with the feature map [32], using the Sigmoid as activation function is commonly used after extracting the attention to represent the distribution between 0 to 1 from the highest value with the following Equation (1):

$$Sigmoid = \frac{1}{1+e^{-x}} \quad (1)$$

To mitigate the problem, we proposed to use residual connection instead of multiplication with sigmoid output as the attention mechanism. We called this approach as ResAtt for the rest of this paper. The output of the ResAtt attention used the ReLU activation function and then summed with the result of the DenseNet block, thus allowing it to increase or decrease the sensor value from the block source residually. An overview of how ResAtt Extension worked can be seen in Figure 4. ResAtt block is implemented on every DenseNet-121 block with each of them producing the same channel as the input. The architecture of ResAtt Densnet as a whole can be seen in Figure 5.

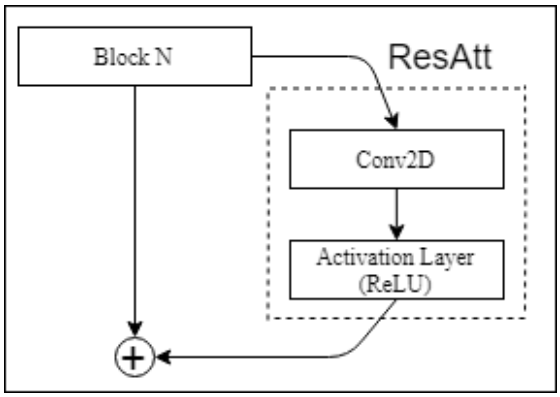


Figure 4: ResAtt DenseNet flow

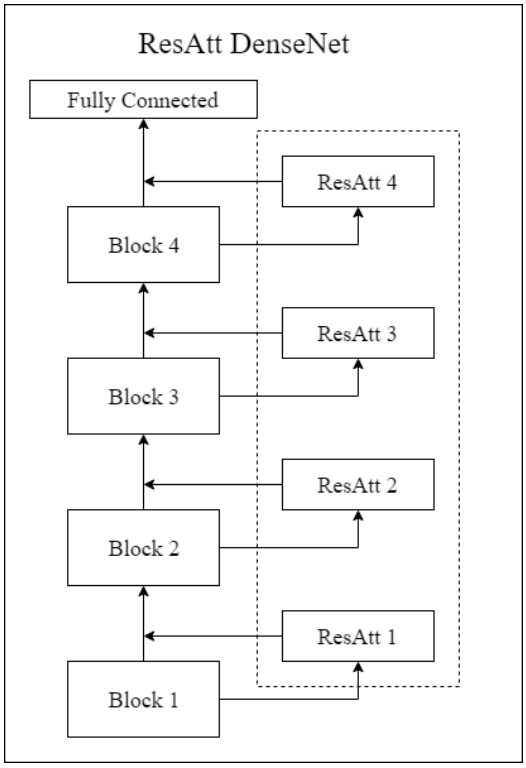


Figure 5: ResAtt DenseNet architecture

IV. Dataset

All of the architectures mentioned were tested on a dataset of oil palm fruit image, each image has been classified by the experts with a total of 400 images and 7 class of ripeness.

Since the number of images is small, a preprocessing method called Ten Crop was used to crop each picture from different sides with horizontal flipping, which became 10 images from a single image. An example of the preprocessing can be seen in Figure 6. There were seven classes of ripeness that are distributed unevenly.

(a) Original image

(b) After Ten Crop preprocess

Figure 6: Dataset preview with Ten Crop preprocess

The dataset was divided into three groups: training data (60%, 253 images), validation (20%, 80 images), and testing (20%, 77 images) from the total images, as tabulated in Table 2. We matched the test split percentage to the Ibrahim et al. study [11]. We also split the dataset for validation, since the use of validation split has become a standard for transfer learning [33]. The percentage of the validation split was matched to the test split percentage.

Table 2
Classification of Oil Palm Fruit

Name	Description	Total
BP	Ripening	16
BM	Raw	8
KM	Less Ripped	64
MKM	Almost Ripped	16
M	Ripped	96
MM	Perfectly Ripped	168
TM	Too Ripped	32
	Total	400

V. Results and Discussion

To evaluate the model's performance, we used accuracy and F1 Score as the evaluation metric. All architectures were implemented with PyTorch under the same python environment on NVIDIA Tesla P100 and Tesla P4 GPU provided by NVIDIA - BINUS AI R&D Center. All models were pretrained on ImageNet (transfer learning), so the models have already learned multiple features from the general objects. The hyperparameters configuration can be seen in Table 3.

Table 3
Hyperparameter Config

Hyperparameter	Value
Learning Rate (LR)	0.001
LR decay steps	Every 8 steps dropped for 10^{-1}
Optimization	SGD (Adam for AlexNet)
Batch Size	8
Epochs	50
Freeze convolutional parameters	False (except AlexNet)

The LR used is the recommended learning rate value by Kingma et al. [34]. The learning steps have been tested both on every 8 and 30 epochs with no significant results changed, therefore 8 epochs were used as the default for every model tested. The optimization method used is Stochastic Gradient Descent (SGD), since it tends to work better on similar data between classes. The convolutional parameters of all models except AlexNet were not frozen so that the network could continue learning new features on top of pretrained from ImageNet. For AlexNet, the experiment shows that it learns better if the convolutional parameters were frozen.

The lowest validation loss and accuracy of all models are reported in Table 4. The lowest validation loss and the highest accuracy was achieved by ResAtt Densenet, showing the superior performance of our proposed attention mechanism module. The plot of validation losses at each epoch for all models is given in Figure 7.

Table 4
Training Results

Model Name	Lowest Validation Loss	Accuracy
AlexNet	0.9829	0.60
DenseNet + SE layer	0.8500	0.64
ResAtt DenseNet	0.4641	0.69
DenseNet Sigmoid	0.8864	0.69

To understand the behavior of ResAtt DenseNet, we provide samples of the generated attention for each DenseNet block in Table 5. The attention samples were generated by using a sample of images for each class.

From Table 5, a conclusion can be made that the block 1 attention focuses on distinguishing the fruit on the palm or from the background image. As the block goes on, the model concentrates on a specific pattern or feature from the image. For example, on block 4 attention for M class, the model focuses on the red part of the palm fruit. However, the model could sometimes focus on unimportant areas from a human perspective, as shown in block 2 attention for MKM class.

The F1 Scores of each model are shown in Table 6, ResAtt DenseNet has been proved to have the highest performance among all of the tested models.

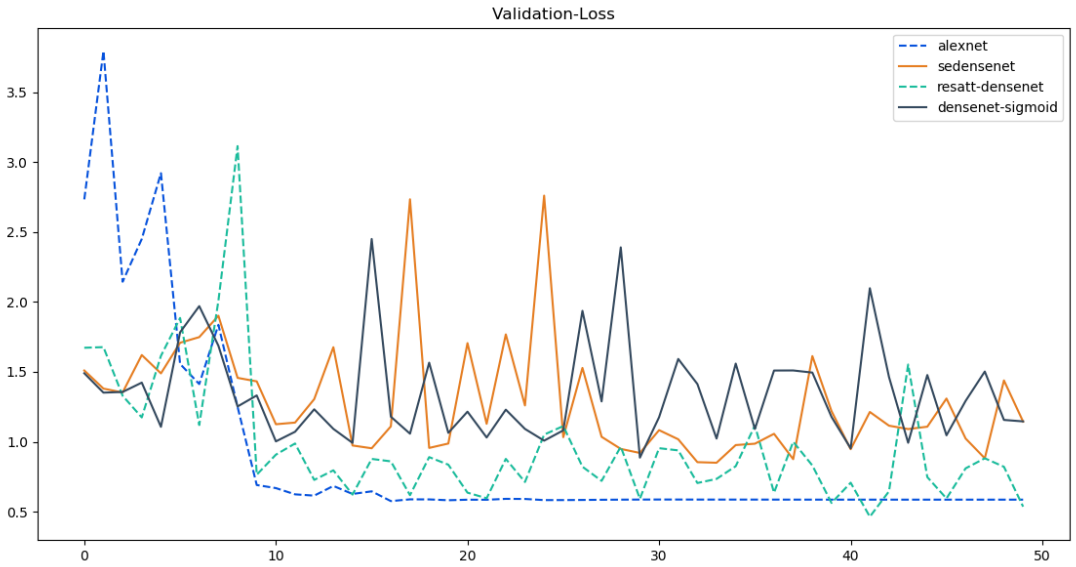
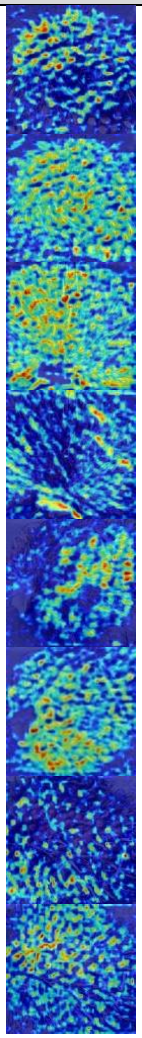
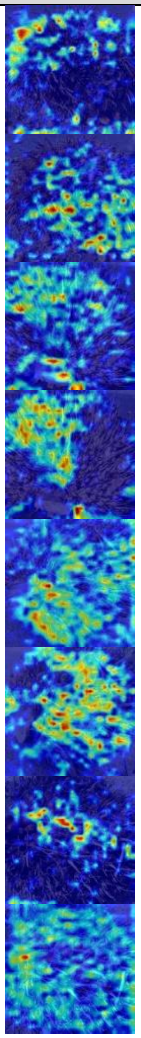
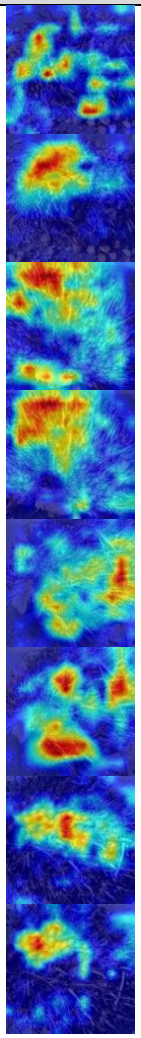
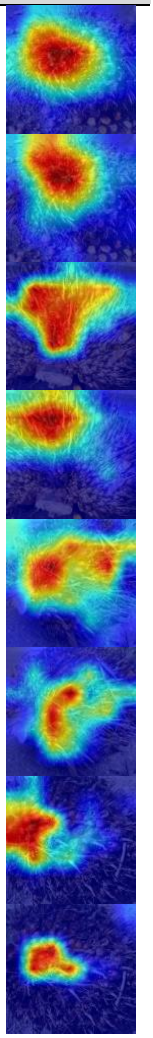


Figure 7: All proposed models validation loss during training

Table 5
ResAtt DenseNet Attention Visualization

Class	Original Image	Block 1 Attention	Block 2 Attention	Block 3 Attention	Block 4 Attention
BP					
BM					
KM					
M					

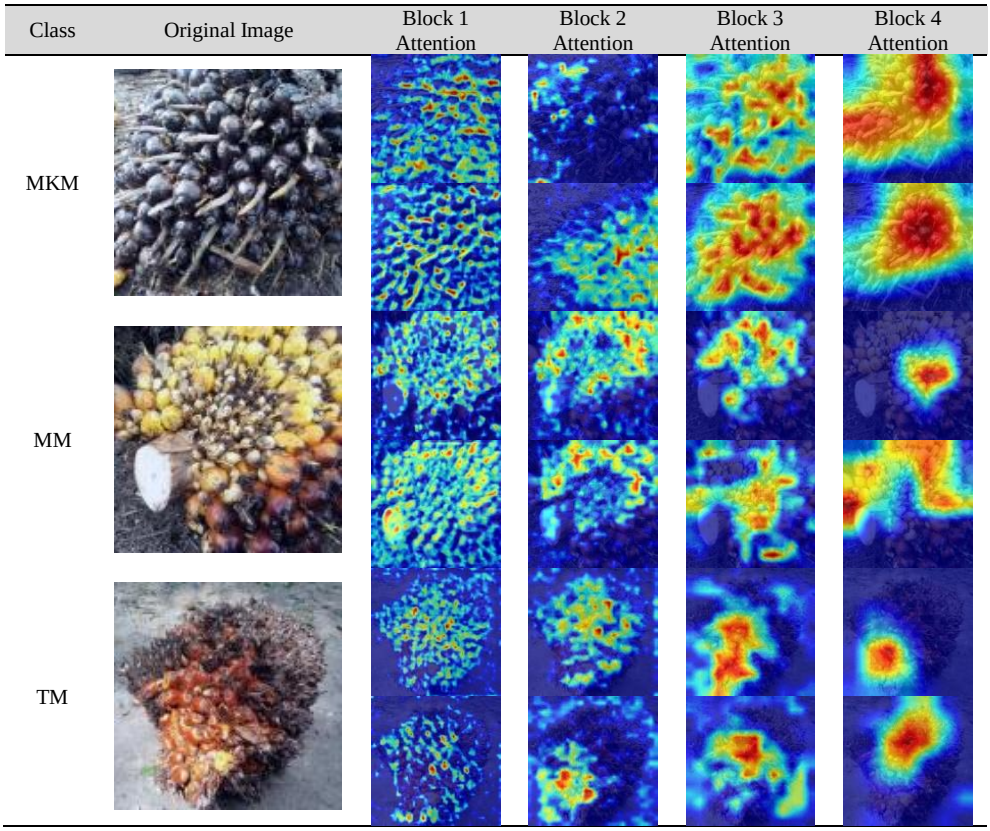


Table 6
Test F1 Score

Model Name	Precision	Recall	F1 Score
AlexNet	0.5951	0.5974	0.5932
DenseNet + SE Layer	0.6458	0.6364	0.6223
ResAtt DenseNet	0.7152	0.6883	0.6929
DenseNet Sigmoid	0.7087	0.6883	0.6819

Table 7
ResAtt DenseNet Confusion Matrix

		Prediction							
		BP	BM	KM	MKM	M	MM	TM	Total
Ground Truth	BP	3	-	-	-	-	-	-	3
	BM	-	1	-	-	-	-	-	1
	KM	-	-	7	1	-	4	-	12
	MKM	-	1	1	1	-	-	-	3
	M	-	-	-	-	16	1	2	19
	MM	-	-	9	-	3	21	-	33
	TM	-	-	-	-	2	-	4	6
	Total	3	2	17	2	21	26	6	77

In addition, we provide a confusion matrix to show a more detailed performance comparison between ResAtt and other attention-based models in Table 7, 8, and 9. The green-colored cells are the best performance of the corresponding class. These confusion matrices show that the ResAtt DenseNet achieved the best performance on three out of seven classes, which is the same as the DenseNet with Sigmoid attention. However, the ResAtt Densenet is still superior to the DenseNet with Sigmoid attention model, because it achieved a better F1 score. On the other hand, DenseNet + SE Layer could only achieve the best performance on two out of seven classes.

Table 8
DenseNet + SE Layer Confusion Matrix

		Prediction							
		BP	BM	KM	MKM	M	MM	TM	Total
Ground Truth	BP	-	-	-	-	1	-	2	3
	BM	-	-	-	1	-	-	-	1
	KM	-	-	8	-	-	4	-	12
	MKM	-	-	1	2	-	-	-	3
	M	-	-	2	-	17	-	-	19
	MM	-	-	9	-	4	20	-	33
	TM	-	-	-	-	4	-	2	6
	Total	-	-	20	3	26	24	4	77

Table 9
DenseNet Sigmoid Attention Confusion Matrix

		Prediction							
		BP	BM	KM	MKM	M	MM	TM	Total
Ground Truth	BP	3	-	-	-	-	-	-	3
	BM	-	-	-	1	-	-	-	1
	KM	-	-	5	2	2	3	-	12
	MKM	-	-	-	3	-	-	-	3
	M	-	-	1	-	16	1	1	19
	MM	-	-	2	-	7	23	1	33
	TM	-	-	-	-	3	-	3	6
	Total	3	-	8	6	28	27	5	77

VI. Conclusion

The attention generated residually from neural networks has been proved useful as an extension to generate attention by convoluting the result of each block on a deep neural network such as DenseNet. It also achieved a better performance than its counterpart that used Sigmoid attention,

which proved its ability in reducing vanishing gradient problem for attention module. Future works can explore the use of other attention mechanism strategies that can cope with vanishing gradient problem, such as extra losses after.

ACKNOWLEDGMENT

The authors are deeply thankful and would like to acknowledge the BINUS AI R&D Center for granting permission to conduct and guiding the research as well as for the resources (NVIDIA Tesla P100 and Tesla P4).

REFERENCES

[1] T. W. Cenggoro, A. Budiarto, R. Rahutomo, and B. Pardamean, "Information System Design for Deep Learning Based Plant Counting Automation," *1st 2018 Indones. Assoc. Pattern Recognit. Int. Conf. Ina. 2018 - Proc.*, pp. 329–332, 2019.

[2] R. Rahutomo, A. S. Perbangsa, Y. Lie, T. W. Cenggoro, and B. Pardamean, "Artificial Intelligence Model Implementation in Web-Based Application for Pineapple Object Counting," in *2019 International Conference on Information Management and Technology (ICIMTech)*, 2019, no. August, pp. 525–530.

[3] "Palm Oil Exports By Country," *Index Mundi*, 2014. [Online]. Available: http://www.indexmundi.com/agriculture/?commodity=palm-oil&graph=exports.

[4] O. M. Bensaeed, A. M. Shariff, A. B. Mahmud, H. Shafri, and M. Alfatni, "Oil palm fruit grading using a hyperspectral device and machine learning algorithm," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 20, no. 1, 2014.

[5] M. S. M. Alfatni, A. R. M. Shariff, H. Z. M. Shafri, O. M. Ben Saaed, and O. M. Eshanta, "Oil Palm Fruit Bunch Grading System Using Red, Green and Blue Digital Number," *Journal of Applied Sciences*, vol. 8, no. 8. pp. 1444–1452, 2008.

[6] N. Jamil, A. Mohamed, and S. Abdullah, "Automated grading of palm oil Fresh Fruit Bunches (FFB) using neuro-fuzzy technique," *SoCPaR 2009 - Soft Comput. Pattern Recognit.*, no. July 2015, pp. 245–249, 2009.

[7] N. Fadilah and J. Mohamad-Saleh, "Color feature extraction of oil palm fresh fruit bunch image for ripeness classification," *13th Int. Conf. Appl. Comput. Appl. Comput. Sci.*, pp. 51–55, 2014.

[8] W. I. W. Ishak and R. M. Hudzari, "Image based modeling for oil palm fruit maturity prediction," *J. Food, Agric. Environ.*, vol. 8, no. 2, pp. 469–476, 2010.

[9] Harsawardana *et al.*, "AI-Based Ripeness Grading for Oil Palm Fresh Fruit Bunch in Smart Crane Grabber," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 426, p. 012147, Mar. 2020.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Inf. Process. Syst.*, 2012.

[11] Z. Ibrahim, N. Sabri, and D. Isa, "Palm Oil Fresh Fruit Bunch Ripeness Grading Recognition Using Convolutional Neural Network," *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 3, pp. 109–113, 2018.

[12] Y. LeCunn *et al.*, "Backpropagation Applied to Handwrittern Zip Code Recognition," *Neural Comput.*, vol. 1, pp. 541–551, 1989.

[13] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009.

[14] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 770–778, 2016.

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[17] K. K. Evans, "Visual Attention," *Wiley interdiciplinary Rev. Cogn. Sci.*, vol. 2, no. 5, pp. 503–514, 2011.

[18] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, Feb. 2015.

[19] S. M. A. Eslami, N. Heess, T. Weber, Y. Tassa, K. Kavukcuoglu, and G. E. Hinton, "Attend, Infer, Repeat: Fast Scene Understanding with Generative Models," 2016.

[20] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning Context Flexible Attention Model for Long-term Visual Place Recognition," *IEEE Robot. Autom. Lett.*, vol. PP, no. c, pp. 1–1, 2018.

[21] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," *Adv. Sp. Res.*, vol. 55, no. 11, pp. 2493–2499, Jun. 2015.

[22] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to Scale: Scale-aware Semantic Image Segmentation," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 3640–3649, 2016.

[23] Q. Xiao, G. Li, L. Xie, and Q. Chen, "Real-world plant species identification based on deep convolutional neural networks and visual attention," *Ecol. Inform.*, vol. 48, pp. 117–124, 2018.

[24] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 936–944, 2017.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 7132–7141, 2018.

[26] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, no. November 2014, pp. 447–456, 2015.

[27] C. Couprie, L. Najman, and Y. Lecun, "Learning Hierarchical Features for Scene Labeling," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 35, no. 8, pp. 1915–1929, 2013.

[28] A. Krizhevsky and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[29] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*, 2018, pp. 103–110.

[30] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," 2017.

[31] B. Pardamean, T. W. Cenggoro, R. Rahutomo, A. Budiarto, and E. K. Karuppiah, "Transfer Learning from Chest X-Ray Pre-trained Convolutional Neural Network for Learning Mammogram Data," *Procedia Comput. Sci.*, vol. 135, pp. 400–407, 2018.

[32] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to Scale: Scale-Aware Semantic Image Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3640–3649.

[33] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 2656–2666, 2019.

[34] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *The International Conference on Learning Representations 2015*, 2015.