Hindawi
Scientific Programming
Volume 2023, Article ID 4301632, 10 pages
https://doi.org/10.1155/2023/4301632

Research Article

Is Vehicle Plate Corner Prediction by Vision Transformer Better than CNNs?

Kyungkoo Jun [1,2]

[1]*Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, Republic of Korea*
[2]*Energy Excellence and Smart City Lab, Incheon Nation University, Incheon, Republic of Korea*

Correspondence should be addressed to Kyungkoo Jun; kjun@inu.ac.kr

Received 13 June 2022; Revised 6 December 2022; Accepted 6 January 2023; Published 27 January 2023

Academic Editor: Roberto Natella

The license plate recognition performance can be improved by converting the license plate photographed from the side to the front view. To perform this transformation, four vertex corner positions of the license plate are required. Existing deep learning methods to find these corner positions use a convolutional neural network (CNN). In this study, we propose a model using a vision transformer (ViT), a nonconvolutional method that has recently been attracting attention, as a backbone, and compare its performance with existing CNN models. The ablation study is conducted by diversifying the ViT structure. Through these results, it was found that ViT has strengths in model size reduction but is similar in performance or inferior to CNN and that ViT training is more difficult than CNN.

1. Introduction

License plate recognition (LPR) generally consists of two steps. In step 1, the license plate area is found within the image, and in step 2, character recognition is performed for the area. If the license plate in the area is not photographed from the front, as shown in Figure 1, it negatively affects the character recognition. Therefore, in order to improve recognition performance, it is necessary to rectify the tilted license plate into the form viewed from the front. This can be done through a perspective transform, which is a warp operation that maps the coordinates of the four corners of the license plate to the positions of the rectangular corners.

The rectification not only improves recognition performance but also increases the efficiency of labeling work that creates training data. For example, it is easier to mark rectangular letters than slanted letters with rectangular bounding boxes. Also, it facilitates the training process of the deep models for LPR. With rectified plate images, the training can lower the dependence on augmentation to consider tilting. For this reason, in the field of document character recognition, it is a useful technique because it can increase accuracy by unwarping inclined or nonplanar document images.

However, the rectification task is challenging due to the real-time requirements of LPR and the limitations of available computing resources. When performing LPR, the computing resources are insufficient because the operation is performed on a low-profile device directly connected to the camera. Also, it should be considered that rectification is a secondary task that must be performed with a minimum of resources so that time and resources can be sufficiently used in the process of locating the license plate area and character recognition.

In this study, we develop a model to detect corner positions for vehicle plate rectification using a vision transformer (ViT), which has recently been attracting attention as a nonconvolutional model, as a backbone, and compare the performance with existing CNN-based methods. ViT was first applied to the vision classification field [1], and its application has been expanding to the detection and segmentation field [2]. ViT is also versatile in its composition and is also used in hybrid form when combined with other well-known models [3].

However, recently, there have been studies that question the claim that ViT performs better and is more robust than CNN [4]. In order to broaden our understanding of the pros and cons of ViT, we intend to create 600 corner prediction models through a combination of structural parameters and analyze the effect of parameters through performance measurement. The contributions of this study can be summarized as follows:

(i) We developed a ViT-based corner prediction model that can be used for the rectification of tilted license plates in images

(ii) Through an ablation study on 600 models, the effect of structural parameters on model size and performance was analyzed

(iii) By comparing the performance of ViT-based models with CNN-based ResNet and MobileNet, the advantages and disadvantages of ViT were analyzed

This article is organized as follows: Section 2 examines license plate rectification-related studies and vision tasks using ViT, and Section 3 proposes a ViT-based plate corner prediction model. In Section 4, performance evaluation is carried out, and Section 5 concludes the study.

2. Related Works

Finding the four corner positions for the rectification of the tilted license plate image can be thought of as a kind of detection problem. If the outline of the license plate is clear, edge detection can be used. However, since there are many images with unclear outlines, there is a limit to its general application. As another method, we can think of a method to find the outermost line that includes all the letters, but it is not efficient because we have to detect the letters first.

Methods [5, 6] using CNN for corner prediction use features extracted from convolutional layers or predict corner coordinates from a latent representation created by an autoencoder. While these methods find the coordinates required for rectification through warping, a method for directly creating a rectified image [7] has also been proposed. For example, there is a method using U-net [8]. However, this method is difficult to apply to the license plate recognition task with real-time requirements. This is because denoising is required as a preliminary step to increase the quality of the rectified image, and the image creation process takes more time than warping. Also, there are cases where blurring occurs in the resulting image, which leads to a decrease in recognition performance [9]. A domain adaptation technique was tried in [10], where a prediction model trained for the plates of one country is adapted to the same task but for the plates of a different country.

Recently, ViT, which is expected to replace CNN in various vision tasks, was based on a representative transformer model [11] used in the field of natural language processing. ViT converts the input image into a sequence of image patches and extracts context information through a self-attention structure with multiheads. This information is evenly distributed between layers and is known to be more advantageous for maintaining spatial information [12]. In addition, it is more robust to adversarial perturbations as well as occlusion and domain shift compared to the existing convolution-based structures [13, 14]. As an example of ViT being used for rectification, it was used to unwarp document images with geometric distortion [15]. Here, the transformer corrects the distortion through pixel-wise displacement.

The self-attention mechanism, a core element of ViT, is limited to low-resolution image input due to computational complexity, and there have been questions about whether it will be effective for tasks that require high input resolution, such as detection or segmentation. However, for the detection task, a model has been proposed that uses ViT as a backbone and combines a common detection task head [16]. In the segmentation task field, it is showing better performance than CNN in the form combined with the existing U-Net structure [17]. Image generation using generative adversarial networks is known to be the most difficult of vision tasks. Even in this field, competitive results were obtained using only ViT without the help of CNN [18]. Recently, it has been believed that ViT will completely replace the convolution operator, which was considered essential for vision tasks. Contrary to this belief, CNNs are favorably evaluated by claiming that convolution-based networks can have as much adversarial robustness as transformers if they follow a learning methodology similar to that of transformers [4].

The reason we considered ViT instead of CNN as a backbone is to test whether the ViT backbone can show competitive performance while having a smaller size than CNN. As described above, since rectification is a secondary process in the license plate recognition process, it should use as few resources as possible, and it should be possible to omit it if necessary. This is important because most license plate recognition processes are performed on edge devices with limited computing resources. In such a low-profile environment, a new model that is smaller than the convolution-based model but can perform comparably to it is needed. So, we want to judge the potential of ViT as a backbone, which has recently been receiving attention.

3. Plate Corner Prediction by Vision Transformer

The proposed model for plate corner prediction consists of a ViT backbone and corner head, as shown in Figure 2. ViT takes the input image as a sequence of patches and generates encoded patches through a self-attention mechanism. The corner head consists of a fully connected layer that predicts 8 values corresponding to 4 corner coordinates (x, y).

The lower part padding is applied to the input image, as shown in Figure 1, in order to include the entire license plate area within the square shape. As a result, in most input images, the actual license plate occupies the top portion of the square image. The license plates in the training and test images are of two types: those for cars and those for motorbikes. The license plate for automobiles in Figure 1(a) is rectangular, and the license plate for motorbikes in Figure 1(b) has a complex shape with the upper part narrower than the lower part.

(a) (b)

Figure 1: Examples of tilted vehicle plates (a) and motorbikes (b).

The input image is divided into patches of size $P_{sz} \times P_{sz}$, as used in the existing ViT. These patches are composed of a sequence of image patches according to the row-major order and are converted into tokens through the patch embedding process. Position encoding is added to each patch and provided as an input to a transformer encoder where self-attention operation is performed. ViT for the classification task uses a separate class token, but it is not used in our proposed model. This is because corner prediction is similar to a detection task rather than a classification task, so all tokens, including spatial information about the image, must be used.

Figure 3 shows the positions in the license plate image of the four coordinates that the corner head should predict. The eight values predicted by the model are x and y coordinate values between 0 and 1, which are normalized according to the image size. As shown in Figure 3(a), each number is associated with one of the four corner positions, such as top-left, top-right, bottom-left, and bottom-right. This matching relationship is equally applicable in the case of a motorbike plates, as shown in Figure 3(b). However, since it is not in a rectangular shape, the upper corner positions are determined so that rectification is better achieved when perspective transformation is performed. For example, the top corner positions are not located on the top-most part of the license plate but rather where the rectangular shape starts.

As summarized in Table 1, the proposed ViT model can have different configurations by changing the combination of four parameters: patch size P_{sz}, depth D_{attn}, number of multi-heads H_{num}, and embedding dimension, E_{dm}. Regarding P_{sz}, input 2D image $I \in R^{H \times W \times C}$ is spliced into a sequence of 2D patches $I_p \in R^{N \times P_{sz} \times P_{sz} \times c}$, where (H, W) is the resolution of the original image, C is the number of channels, (P_{sz}, P_{sz}) is the resolution of each image path, and $N = HW/P_{sz}^2$. D_{attn} is the number of alternating layers of self-attention, vertically stacked inside the transformer encoder. H_{num} is the number of heads in a multiheaded self-attention module. E_{dm} is constant latent vector size by which image patches I_p are mapped to E_{dm} dimensions through a trainable linear projection. The vector size is also called the token length.

The values of E_{dm} is constrained by H_{num} because the embedded vector size E_{dm} must be divisible by H_{num}, satisfying the following equation:

$$E_{dm} \geq k \cdot H_{\text{num}}, \tag{1}$$

where $k > 1$, a positive integer. To keep computing and number of parameters evenly distributed over self-attention modules when changing H_{num}, E_{dm} is typically set to multiples of H_{num}.

To this end, given the input image resolution of 416×416, which are used in our experiments, Table 1 shows all of possible 600 configurations made from the four parameter combinations. For example, the parameters can have the following values: $P_{sz} \in \{13, 26, 52, 104, 208\}$, $D_{attn} \in \{4, 8, 16, 32, 64, 128\}$, and $H_{\text{num}} \in \{2, 4, 8, 16, 32\}$. The reason that E_{dm} is determined according to H_{num} is that when creating a multihead structure, the length of the token input to each head is set to E_{dm}/H_{num}; when $H_{\text{num}} = 2$, $E_{dm} \in \{4, 8, 16, 32, 64, 128\}$, whereas when $H_{\text{num}} = 32$, $E_{dm} \in \{64, 128\}$.

To understand the effect of these structural parameters on the sizes of the proposed ViT model, the parameters' correlations with the number of model parameters were analyzed for all of 600 models. As shown in Figure 4(a), E_{dm} has the strongest positive correlation with number of model parameters, followed by D_{attn} and H_{num}, and P_{sz} has the weakest correlation. Because we flatten image patches with size of $P_{sz} \times P_{sz}$ and map them to E_{dm} dimensions through convolutional projection, the number of parameters directly related to P_{sz} is not significant. On the contrary, the embedded tokens having the length of E_{dm} are passed through all layers, the amount of related parameters proportionally increases as E_{dm} grows. D_{attn} represents the number of vertically stacked layers of self-attention, thus contributing to the increase of the model parameters. The same explanation applies to H_{num}. We also analyze the correlation between the four parameters. We observe that there exists positive correlation only between E_{dm} and H_{num}. Due to the structural characteristics of the ViT model, as we increase E_{dm}, the possible value range of H_{num} widens, as in (1), resulting in increased model size.

Figure 4(b) shows the number of model parameters according to the values of E_{dm} and D_{attn}, the two elements with the strongest positive correlation. When $E_{dm} = 128$ and $D_{attn} = 128$, the model size is the largest, and the number of model parameters is 30 million, which is greater than the 23.0 million of ResNet-50 and less than the 42.6 million of ResNet-101. We measure the

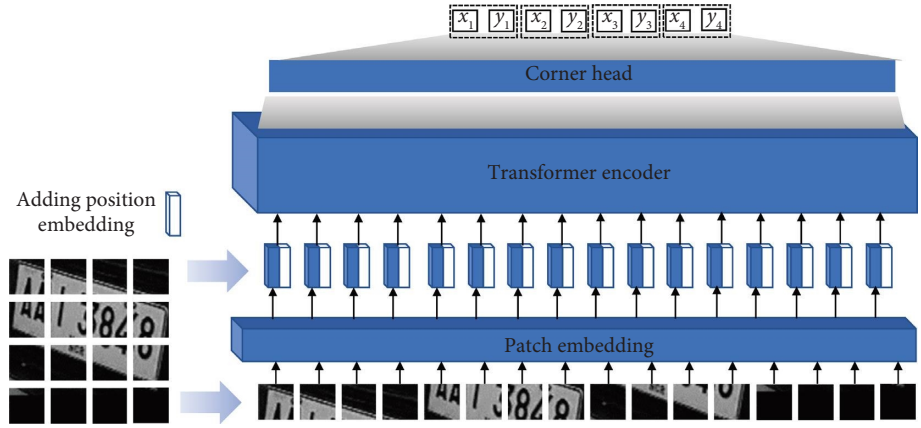


FIGURE 2: The proposed architecture for corner prediction based on vision transformer.

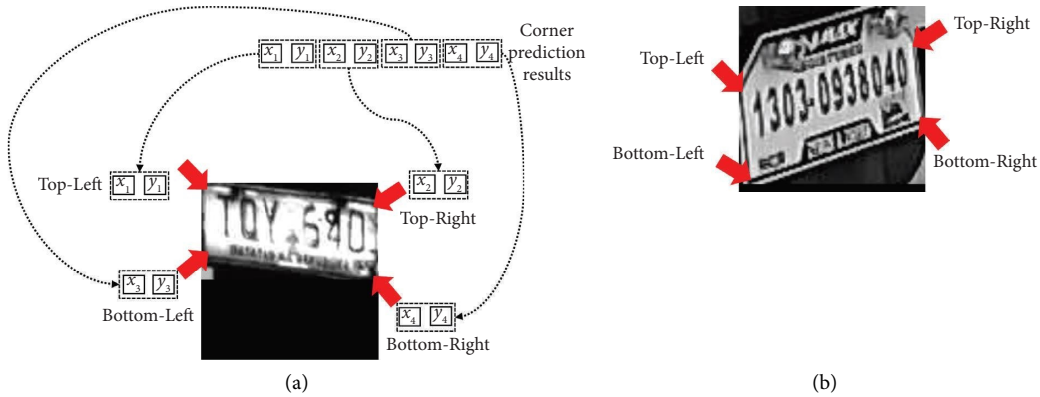


FIGURE 3: Corner positions in plates and the matching relationship from predicted values. (a) Four corner positions in a Philippines plate; (b) corresponding four corner positions in a Philippines motorbike plate.

performance of 600 ViT models corresponding to these combinations using the same data and training under the same conditions, and the results are described in detail in Section 4.

4. Performance Analysis and Ablation Study

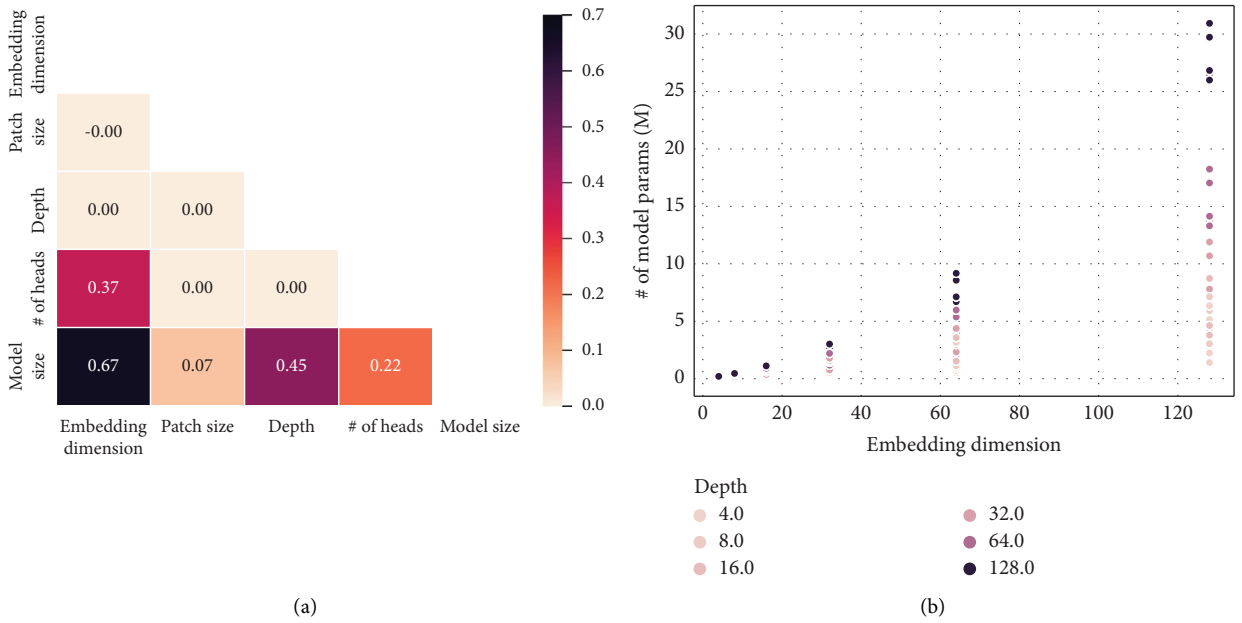
The dataset for model training and performance measurement consists of 8,530 training images and 2,134 test images, and all images are in 1-channel gray format. There are two types of plates in the dataset images, one for vehicle plates and one for motorbike plates. The number of motorbike plate images in the training dataset and test dataset is 1,154 and 284, respectively, accounting for about 13% of each dataset. The loss for training the model uses the mean squared error (MSE) between the predicted corner position coordinates and the ground truth. In order to test the generalization of the distribution of the trained model, the performance was measured using 1,000 Korean license plates instead of the Philippines license plates used for training.

Figure 5 shows the loss distribution of the validation dataset after training all 600 ViT models with the same training data for epoch = 50. Figure 5(a) shows the distribution of loss values according to patch size P_{sz}. When it is the smallest patch size, with $P_{sz} = 13$, the loss distribution range is the widest, and the loss values are generally larger than those of other patch sizes. When the patch size is small, it is difficult to achieve good performance because the spatial context of the image cannot be sufficiently contained. Figure 5(b) shows the loss distribution according to the depth D_{attn}. As D_{attn} increases, the average loss also decreases. This is the same reason that performance generally increases as the number of layers included in the deep model increases. Figure 5(c) is the loss distribution according to the embedding dimension E_{dm}. When it has the smallest dimension length with $E_{dm} = 4$, the overall loss value is the largest and the distribution is the widest. This is because, similar to the reason presented in the case of P_{sz}, the embedding tensor is too short to contain enough information. It should be noted that the loss value decreases as it increases up to $E_{dm} = 32$, and then stagnates beyond that, which means that the performance improvement by increasing the dimension is limited. In Figure 5(d), it can be seen that the loss does not change significantly according to the number of heads H_{num}, which means that H_{num} does not directly affect the performance.

The results of Figure 5 helps understand the ViT model characteristics in the theoretical analysis aspects based on the discussion of Section 3. Unlike prior works using self-

Table 1: Different configurations of the proposed ViT model, given the input image resolution of 416 × 416 according to combinations of four parameters: patch size P_{sz}, depth D_{attn}, number of multiheads H_{num}, and embedding dimension E_{dm}. In our experiments, a total of 600 possible configurations can be generated under configuration constraints.

Patch size (P_{sz})	Depth (D_{attn})	Number of multiheads (H_{num})	Embedding dimension (E_{dm})
13, 26, 52, 104, 208	4, 8, 16, 32, 64, 128	2	4, 8, 16, 32, 64, 128
		4	8,16, 32, 64, 128
		8	16, 32, 64, 128
		16	32, 64, 128
		32	64, 128



(a) (b)

Figure 4: Correlation analysis of the ViT model size with the four structural parameters (a) and the number of model parameters according to varying E_{dm} embedding dimension and D_{attn} depths (b).

attention in computer vision, we do not introduce task-specific inductive biases into the proposed model architecture apart from the initial image patch extraction step. Instead, we interpret the ViT model as a general and scalable structure that can be configured by adjusting the four key structural parameters. Such structural exploration by the combination of the parameters allows the ViT model to integrate critical regression information across the entire image even in the relatively initial stages, with the help of self-attention. Specifically, we observe the most structurally profound impact on performance can be realized by D_{attn} the depth, which most exploits the advantages of self-attention. Also, H_{num} the number of multiheads with self-attention that implement attention distance is analogous to the receptive field size in CNNs, resulting in linearly increasing accuracy as H_{num} increases.

From the validation loss results, we selected one model with the best performance for each P_{sz}, a total of 5 models. Table 2 shows the configuration parameter values and number of model parameters of these models. For easy identification of these models, they are named according to P_{sz}. The smallest size model is ViT-Patch-13, which has less than 1 million parameters. The largest model is ViT-Patch-104, with 26.8 million parameters. All models were configured to have a sufficiently deep depth with $D_{attn} \geq 64$.

ResNet [19] and MobileNet [20, 21], which show excellent results in various vision tasks, are used as the backbones of CNN-based models for the performance comparison with the proposed ViT model. The reason that we choose ResNet for comparison is that it was typically in the original work [1] that proposed the vision transformer, ResNet was used as the baseline CNNs to compare the feature extraction capability of backbones. We consider MobileNet for a similar reason. In the original work, a variant of EfficientNet [22] was selected as a comparison example of state-of-the-art CNN. Since MobileNet shares similar structural characteristics with EfficientNet but with relatively small sizes, we consider it a proper example to compare the proposed model in the similar experimental setup of the original work. In addition, the size differences between MobileNet and the proposed model provide another comparison metric.

ResNet shows high performance in most vision-related tasks by using residual connections and can make models of various sizes by adjusting the depth. MobileNet is a small

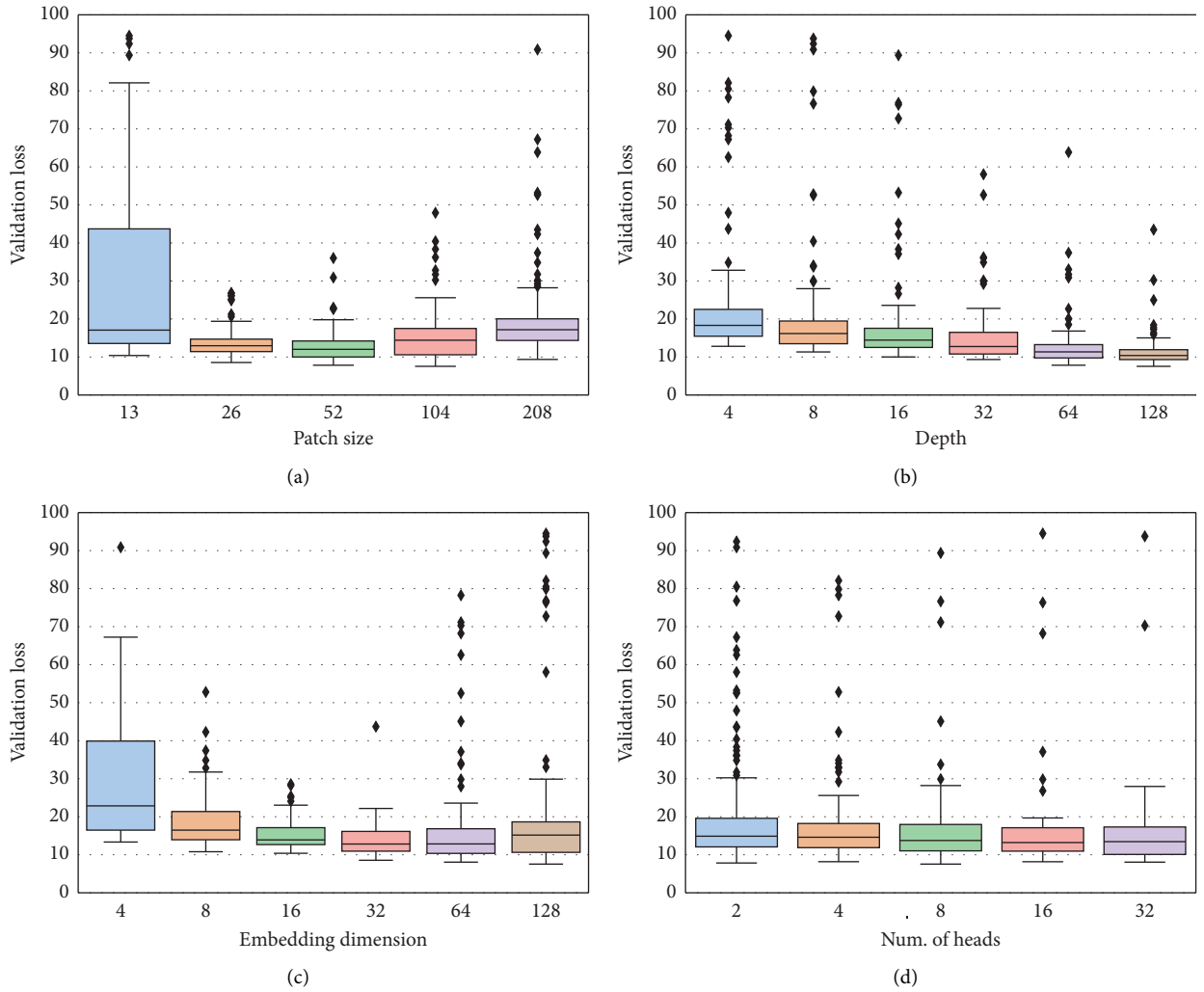


Figure 5: The validation losses of the ViT models according to different combinations of the four structural parameters. (a) Validation loss according to patch size, (b) loss according to depth, (c) loss depending on embedding dimension, and (d) loss in relation with number of heads of the proposed architecture.

Table 2: The number of model parameters of selected ViT backbone models along with corresponding structural parameters.

Model name	Patch size	Embedded dim	Depth	# of heads	# of param (m)
ViT-Patch-13	13	16	128	2	0.96
ViT-Patch-26	26	32	128	2	1.9
ViT-Patch-52	52	128	64	2	13.3
ViT-Patch-104	104	128	128	8	26.8
ViT-Patch-208	208	64	128	8	9.1

size model that is the basis of EfficientNet, which is often used as the backbone for vision tasks and shows good performance thanks to an inverted residual connection with a linear bottleneck. For performance measurement, CNN-based corner prediction models using ResNet 18, 34, 50, 101, 152, and MobileNet V2, V3-Small, and V3-Large as backbones were constructed. Features extracted from these backbones are given as inputs to the same corner head used in the ViT model. That is, the feature contexts extracted from the backbone are converted into 8 values normalized between 0 and 1 through the corner head.

Figure 6 shows the number of parameters of these CNN-based models and the proposed ViT-based models, and these values are directly related to the size of the model. The size of ResNet-based models is relatively large, and MobileNet is generally small, whereas ViT models have various sizes depending on the configuration. These models were trained for 50 epochs with the same data, and in this process, an Adam optimizer set with learning rate $= 1e-4$ and weight decay $= 5e-5$ was used. For the trained models, tests were performed on the images of the Philippines license plate and the Korean license plate, respectively.

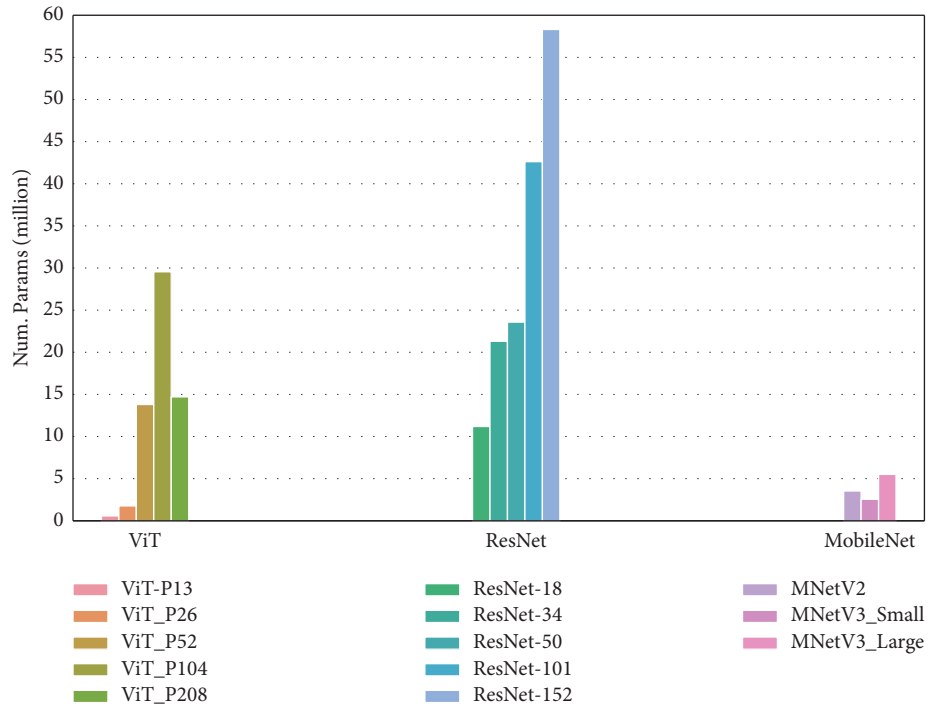


Figure 6: The comparison of the number of model parameters of CNN-based models of ResNet and MobileNet with the proposed ViT models.

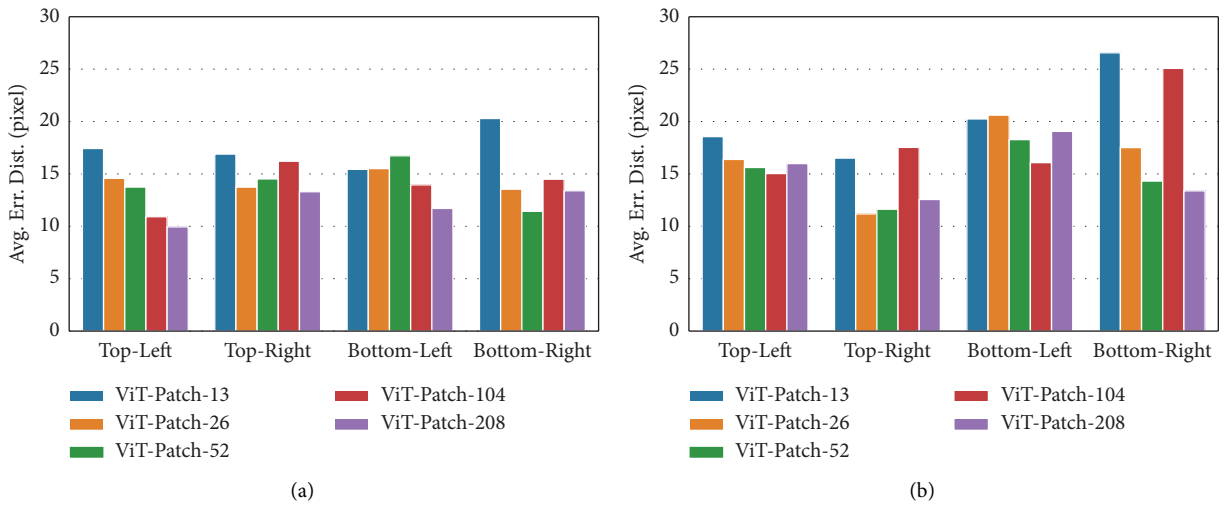


Figure 7: Continued.

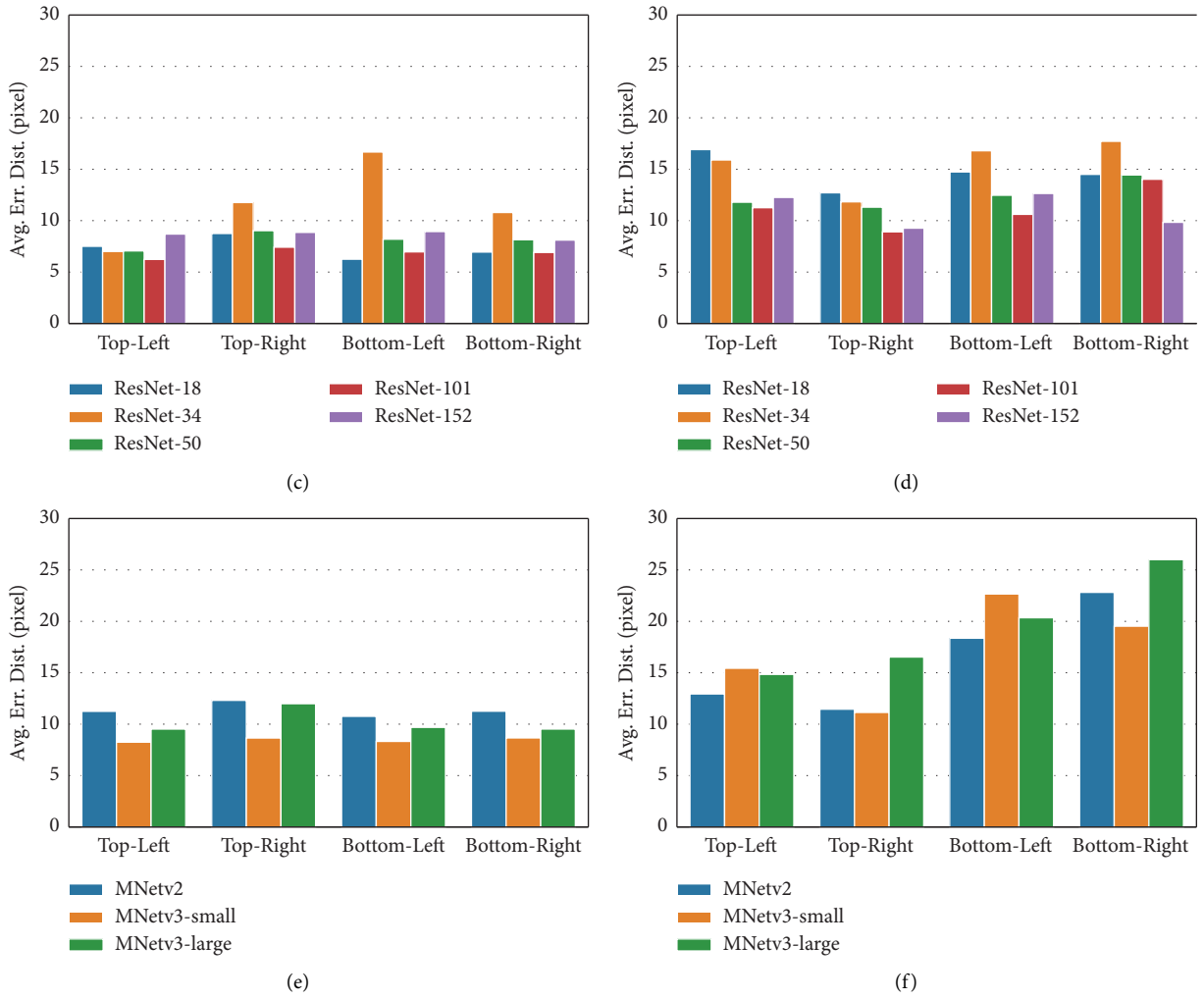


Figure 7: Average prediction errors according to corner positions of the ViT, ResNet, and MobileNet models. (a) ViT error distances for Philippines plates. (b) ViT error distances for Korean plates. (c) ResNet error distances for Philippines plates. (d) ResNet error distances for Korean plates. (e) MobileNet error distances for Philippines plates. (f) MobileNet error distances for Korean plates.

Figure 7shows the average error distance (in pixels) between the predicted corner location and the ground truth for each model for validation data. From the top row of the figure, it corresponds to the results of ViT, ResNet, and MobileNet; on the left is the result for Philippines license plates, and on the right is the result for Korean license plates. And each mean error was displayed as top-left, top-right, bottom-left, and bottom-right according to the corner position.

In Figure 7(a), ViT-Patch-13, the smallest size among ViT models, has a larger error of at least 3~max. 8 pixels than other ViT models. It is noteworthy that, in the case of the second-smaller ViT-Patch-26 model, the error was superior to less than 5 pixels when compared with other larger-sized models. Figure 7(b), the test result for the Korean license plate that is not used for training, has a generally larger error than Figure 7(a), which shows the result for the Philippines license plate, but the error trends for each model are similar. It should be noted here that the largest model, ViT-Patch-104, had a larger error than other models in bottom-right for Korean license plates. Since this model showed adequate performance in the case of the Philippines license plate bottom-right, it is presumed that generalization failed due to model overfit for the Philippines license plate.

In Figure 7(c), ResNet shows an error of about 8 to 12 pixels in most of the 5 models for the Philippines license plate. For Korean license plates, the overall error increased compared to the case of the Philippines, but the larger model showed better performance, and the error was the smallest when compared with other ViT and MobileNet-based models. Figures 7(e) and 7(f) are the results of MobileNet, which show intermediate performance lower than the accuracy of the ResNet-based model and higher than that of the ViT-based model. In the case of the Philippines license plate, the average error is about 10 pixels, and in the case of the Korean license plate, the error distribution is different for each location. For example, lower corner errors are greater

than errors in upper positions. What is unusual is that, unlike in ViT and ResNet, where performance is proportional to model size, the error of MobileNetV3-large with the largest model size is lower or similar to that of the smaller MobileNet-based models.

These results are evidence for two facts, one is that the feature representation ability of CNN is still effective in the corner prediction task. Another is that it is difficult to achieve excellent performance by simply replacing the CNN block with ViT. As in our experiment, when the backbone is replaced from CNN to ViT and undergoes the same training process, ViT shows inferior performance to CNN, suggesting that the use of ViT is more difficult than that of CNN. However, it is noteworthy that the small-sized ViT-Patch-26 model with a number of parameters of 1.9 million showed smaller errors for Korean license plates than the CNN-based models. This means that ViT has potential in terms of generalization considering the model size.

5. Conclusions

As the research results of ViT surpassing existing CNNs in various vision tasks increase, expectations are also growing. The motivation of this study was the question of whether such a dominance is still possible in the license plate corner prediction task. We developed a corner prediction model using ViT as a backbone and compared the performance with existing models using ResNet and MobileNet. As a result, ViT was notable in terms of performance considering size, but ResNet was dominant in absolute performance. These results were obtained through the performance analysis of 600 ViT backbone models created through the combination of four ViT structural determinants.

These results show that performance improvement cannot be achieved by simply replacing the backbone from CNN to ViT. In conclusion, ViT is a more difficult model to handle than CNN; in that, it has to handle the input image more carefully, and the training process is more complicated.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by an INU research grant of 2018-0199.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[2] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: deformable transformers for end-to-end object detection," in *Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, April, 2020.

[3] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3611–3620, Seoul, South Korea, November, 2021.

[4] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are Transformers more robust than CNNs?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 26831–26843, 2021.

[5] H. Yoo and K. Jun, "Deep homography for license plate detection," *Information*, vol. 11, no. 4, p. 221, 2020.

[6] H. Yoo and K. Jun, "Deep corner prediction to rectify tilted license plate images," *Multimedia Systems*, vol. 27, no. 4, pp. 779–786, 2021.

[7] Y. Lee, J. Lee, H. Ahn, and M. Jeon, "SNIDER: single noisy image denoising and rectification for improving license plate recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Seoul, Korea (South), November, 2019.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, palm springs, CA, USA, October, 2015.

[9] X. Huang, Y. Huang, and Y. Pei, "DocGAN: document image unwarping for high-level vision task," in *Proceedings of the IET 8th International Conference on Wireless, Mobile, Multimedia Networks*, Beijing, China, October, 2019.

[10] K. Jun, "Unsupervised domain adaptive corner detection in vehicle plate images," *Sensors*, vol. 22, no. 17, p. 6565, 2022.

[11] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[12] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[13] R. Shao, Z. Shi, J. Yi, P. Chen, and C. Hsieh, "On the adversarial robustness of visual transformers," 2021, https://arxiv.org/abs/2103.15670.

[14] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M. Yang, "Intriguing properties of vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23296–23308, 2021.

[15] H. Feng, Y. Wang, W. Zhou, J. Deng, and H. Li, "DocTr: document image transformer for geometric unwarping and illumination correction," 2021, https://arxiv.org/abs/2110.12942.

[16] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," 2020, https://arxiv.org/abs/2012.09958.

[17] J. Chen, Y. Lu, Q. Yu et al., "Transunet: transformers make strong encoders for medical image segmentation," 2021, https://arxiv.org/abs/2102.04306.

[18] Y. Jiang, S. Chang, and Z. Wang, "Transgan: two pure transformers can make one strong gan, and that can scale up," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14745–14758, 2021.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pp. 770–778, Long Beach, CA, USA, August, 2016.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June, 2018.

[21] A. Howard, M. Sandler, G. Chu et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, Montreal, BC, Canada, October, 2019.

[22] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, Atlanta GA USA, June, 2019.