Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080620
ISSN 2371-8838 Vol. 8 Num. 6

Improved winter road condition classification method for autonomous driving based on hierarchical transformer

Zhengguang Lu[1,a], Huaqi Zhao[1,b,*]

[1]*School of Information and Electronic Technology, Jiamusi University, Jiamusi, China*
[a]*luzhengguang@stu.jmsu.edu.cn*, [b]*huaqijms@163.com*
[*]*Corresponding author*

Keywords: Transformer, Classification, Dynamic overlapping patch embedding, Frequency-based factorized attention, Multi-level feature fusion

Abstract: Aiming at the problem that the existing road surface classification methods cannot accurately identify the conditions of winter road surface with low discrimination, we propose an improved winter road condition classification method based on hierarchical Transformer. Firstly, dynamic overlapping patch embedding is introduced, which can flexibly handle input features of any size and preserve the continuity of local detail information through dynamic position encoding and overlapping patch embedding. Then, frequency-based factorized attention is used to extract frequency features containing high-level context information, enhancing the feature representation between image categories. Finally, a multi-level feature fusion method based on weight average strategy is proposed, by evenly allocating the dynamically upgradable weights to the output layers of each stage and performing multi-level fusion, the low-level features are projected to the high-level to continue learning, the feature representation is enriched, and the discrimination of the classification image is increased, thereby the classification performance is improved. Experiments are carried out on the WRF dataset of snow and ice road. The classification accuracy of the proposed method can reach 88.93% with only 3.8M parameters and 0.6G computational complexity, which is better than the current mainstream road classification methods.

1. Introduction

Image classification tasks provide technical support for the safety of autonomous driving. With the development of computer vision, in order to achieve efficient image classification, some scholars have proposed image classification methods based on deep learning. AlexNet is the first work of convolutional neural network applied to image classification, and its classification performance exceeds other traditional classification methods [1]. Later, Ho proposed that MobileNetv3 improved the classification accuracy by deepening the network depth and residual connection, but they were not effective for some fine-grained classification tasks with small discrimination of image categories [2]. To get rid of the convolutional limitations, Vision Transformer (ViT) utilizes image patches to process local information and model global context information [3]. However, traditional ViT processes images at a single scale. Some people have proposed the Swin Transformer method, which

uses the sliding window to obtain global and detail features [4]. Wang et al. proposed the PVTv2 method with multi-scale hierarchical structure, which uses overlapping cut blocks to maintain local image continuity. The design of pyramid structure is conducive to realizing fine-grained dense classification tasks and improving the classification accuracy of images with low discrimination [5]. In order to further learn discriminative representations of images, some scholars use dynamic position encoding to equip the shallow and deep layers with local and global label affinity, respectively, to achieve efficient representation learning [6]. Due to the computational burden caused by attention in Transformers, some people use adaptive frequency filters to replace the standard self-attention, enhance the category discrimination from the perspective of frequency information, and achieve efficient representation learning with very low computational complexity [7]. Some scholars directly replace the original Transformer module with the pooling layer, and still achieve good classification results [8]. However, these methods have poor classification accuracy for some categories with low discrimination.

In order to ensure that autonomous vehicles can accurately classify the current road state in winter, we propose an improved road state classification method based on hierarchical Transformer. The main innovations include dynamic overlapping position coding, factorized attention based on frequency domain and multi-level feature fusion method based on weight average strategy. Experiments show that this method can effectively improve the classification accuracy.

2. Improved Road condition classification Method based on Hierarchical Transformer

In northern winter, there are often some snow and ice roads with similar pavement conditions characteristics and low recognition, such as little ice, partial snow and a lot of snow. In this road scenario, automatic driving still has the problem of inaccurate classification of road conditions. To this end, we propose an improved road condition classification method based on hierarchical Transformer, and the overall framework is shown in Figure 1. The structure of each stage is similar. Firstly, Dynamic Overlapping Patch Embedding (DOPE) is used to flexibly adapt to different scales of input and maintain the continuity of local detail features. Then, the important frequency features are captured by Frequency-based Factorized Attention (FFA) to enhance the representation between categories. Then, in order to prevent the loss of some details in the process of network learning, a multi-level feature fusion method based on weight average strategy is used to fuse the output features of each stage, and the shallow features are mapped to the high-level full learning to obtain richer feature information. Finally, the classification head outputs the final classification result.

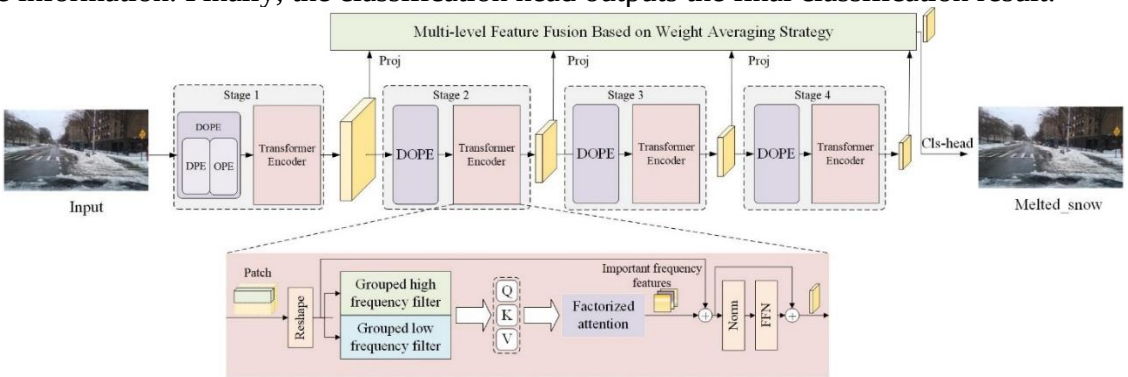

Figure 1: Architecture of improved classification method based on hierarchical Transformer.

2.1. Dynamic overlapping patch embedding

Location information is an important cue to describe visual representations. In order to adapt to

the hierarchical network with different input sizes at each stage and maintain the continuity of local detail information, this paper proposes a dynamic overlapping cut embedding, as shown in Figure 2, which is mainly composed of Dynamic Position Encoding (DPE) based on deep-wise convolution and Overlapping Patch Embedding (OPE). This section will introduce these two parts in detail.

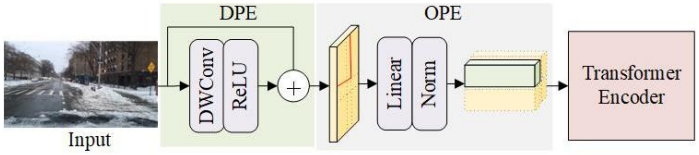

Figure 2: Architecture of dynamic overlapping patch embedding

2.1.1. Dynamic position encoding with deep-wise convolution

Conditional positional encodings implicitly encode positional information through convolution operators, so that transformers can handle inputs of any size. Inspired by this, a dynamic position encoding based on Depth-wise Convolution (DW-Conv) is proposed to flexibly handle inputs of different sizes at different stages, and dynamically integrate position information into all tags. The rationale for choosing DW-Conv is that it can vary the width and height in the spatial dimension without changing the number of channels, which is friendly to arbitrary input shapes; The second is that deep convolutions are lightweight, which is also an important factor in the balance of computational accuracy. Finally, additional zero padding is added that can help a tag learn its absolute position by querying its neighbors step by step. The Dynamic Position Encoding (DPE) is formulated as follows:

$$DPE_i\left(x_i\right)=\sigma\left(DWConv(x_i)\right)+x_i \tag{1}$$

where x_i represents the input feature of stage i $(1 \le i \le 4)$, *DWConv* represents the depth convolution with zero padding, the convolution kernel is *3×3*, the number of input and output channels is equal, and σ represents the *Sigmoid* activation function.

2.1.2. Overlapping patch embedding

Traditional cutting methods divide the image into non-overlapping patches, as shown in Figure 3 (a). This leads to the segmentation of some local features, such as the lumps of snow on the road surface, which can distinguish the road condition, reducing the learning ability of these important features. To this end, this study adopts the method of overlapping patch, as shown in Figure 3 (b). Firstly, zero-padding is used to ensure that the image edge feature information is fully extracted. Then the area of the patch is enlarged so that the adjacent patch can overlap half of the features (that is, the red dashed line in Figure 3 (b)) to maintain the continuity of the local features.

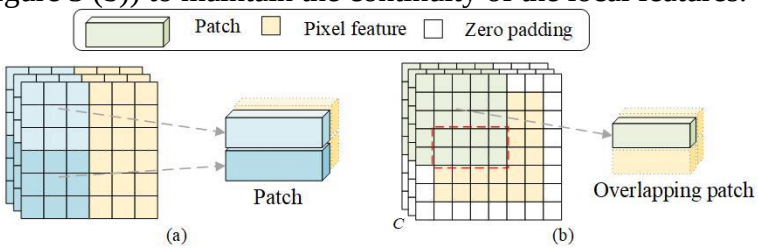

Figure 3: Methods of patch embedding.

Specifically, given an input image size of $H \times W \times C$, let the convolution step size be *S*, the kernel size be *2S-1*, the padding size be *S-1*, and the number of output channels be C', then the output size

is $\frac{H}{S} \times \frac{W}{S} \times C'$. Therefore, the specific process of DOPE is expressed as follows:

$$X_{DOPE} = Conv\left(DPE_i\left(x_i\right)\right) \tag{2}$$

where *Conv* represents obtaining overlapping patches using a *2D* convolution with a step size of 4, a padding of 3, and a kernel of 7. X_{DOPE} represents the final output overlapping patch feature, which is used as the input of the encoder in this level.

2.2. Frequency-based factorized attention

Being able to extract discriminative features from images is of great significance to accurately identify the state categories of winter snow and ice roads. Some scholars believe that frequency information expression can mine the information ignored by human vision and strengthen the differences between learning categories. For the classification of road conditions in winter, we propose frequency-based factorized attention from the perspective of spectral correlation, as shown in Figure 4.

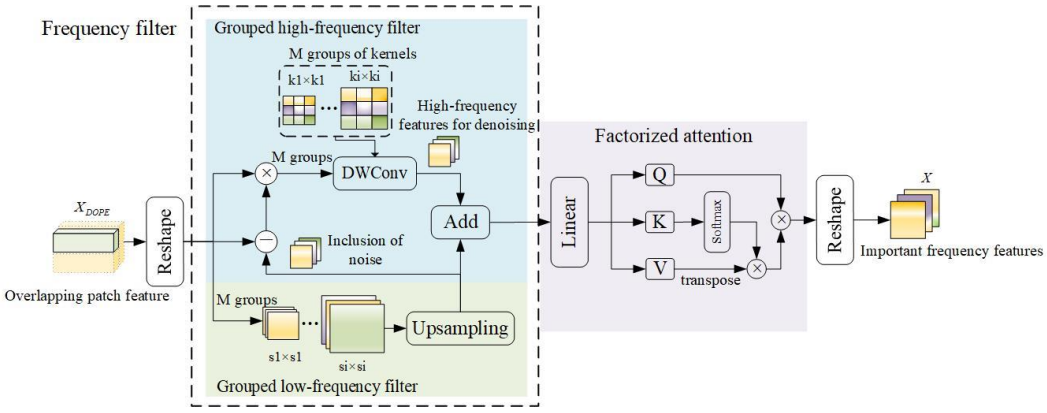

Figure 4: Architecture of frequency-based factorized attention.

2.2.1. Frequency filter

The frequency filter includes two parts: grouped low frequency filter and grouped high frequency filter.

Grouped low frequency filter: The low-frequency information contains a lot of semantic information. The low-frequency filter only allows signals below the cutoff frequency to pass through. The average pooling will be used as the low-frequency filter in this study. Since the cutoff frequency is different for different images, the input is grouped. Assuming m groups, the grouped low-frequency filter is performed as follows:

$$\hat{X} = reshape\left(X_{DOPE}\right) \tag{3}$$

$$GLF^m = Concat\left(BiL\left(\ell_{s\times s}\left(\hat{X}_i\right)\right)\right) \tag{4}$$

where *reshape* stands for dimension transformation, *m* stands for dividing the input into *m* groups, $\hat{X}_i$ stands for the *i*-th group, $\ell_{s\times s}$ stands for adaptive average pooling with output size $s \times s$, *BiL* stands for bilinear interpolation upsampling to ensure that the size of multiple groups of low-frequency features is consistent.

Grouped high frequency filter: High-frequency information can preserve detailed information,

which is crucial in image classification. Convolution can be used as a typical high-frequency operator, which can retain important high-frequency components and filter irrelevant low-frequency components. For each image, the high frequency cutoff frequency is different, and the high frequency component can determine the quality of the image, so here, the high frequency filter is also divided into m groups, and convolution layers with different kernels are used to simulate the cutoff frequency in different high frequency filters and suppress the expression of low frequency noise. The operation details are as follows:

$$GHF^m = Concat\left(BiL\left(Conv_{k\times k}\left(\left(\hat{X} \cdot \left(\hat{X} - GLF^m \right) \right)_i \right) \right) \right) \tag{5}$$

where $Conv_{k\times k}$ represents the depthwise separable convolution with kernel $k\times k$, m groups of denoised high-frequency features are obtained by this method, and finally the denoised high-frequency features are obtained by concatenation.

2.2.2. Factorized attention

In order to obtain the important frequency features with global context information, frequency-based factorized attention is proposed to enhance the expression ability of high and low frequency features. For the input frequency fusion feature $X_f \in R^{H\times W\times C}$, firstly, the key K and value V of the frequency component are calculated through the linear layer. The formula is expressed as follows:

$$X_f = GLF^m + GHF^m \tag{6}$$

$$Q,K,V = reshape\left(Linear\left(X_{LH};W \right) \right) \tag{7}$$

where $Linear$ represents the learnable linear layer, W represents the weight parameters of the linear layer, and reshape Q, K, V into (HW, C). Then, the similarity of frequency features K and V is obtained by normalization and matrix operation. Finally, the important frequency features are selected by querying. The formula is as follows:

$$X = reshape\left(FFA(Q,K,V) \right) = reshape\left(\frac{Q}{\sqrt{d}} \cdot \left(Softmax(K^T) \cdot V \right) \right) \tag{8}$$

where, FFA represents the factorized attention, and X represents the important frequency features of the output. Then, X were processed by residual connection, normalization and feed forward neural Network (FFN) to enhance the local detail features. Let the Transformer encoder depth of the model be N, then the output features of each layer can be expressed as:

$$X_i^\tau = FFN\left(Norm\left(X + \hat{X} \right) \right) + X \tag{9}$$

where $0 \le i < N$, X_i^τrepresents the output features of the Transformer encoder in each stage that contain important frequency information, $Norm$ represents normalization.

2.3. Multi-level feature fusion method based on weight average strategy

In order to prevent the loss of detail information of shallow features in the process of layer-by-layer learning of the model, we propose a multi-level feature fusion method based on weight average strategy, which mainly includes two parts: projection layer and fusion layer, as shown in Figure 5. The purpose of this method is to collect low-level information from shallow features and enrich deep

feature information.

Projection layer: In this study, the projection layer is chosen for the first three stages, specifically, the nonlinear projection is instantiated by the Linear-GELU-Linear structure. Then, the features of the first three stages are reshaped and adaptively pooled, which is guaranteed to be consistent with the output feature dimension of the last layer, so that the output features of the four stages are fused next. Here, only the outputs of S ($S = N - 1$) stages are projected, and the N-th ($N = 4$) output feature X_{N-1}^{τ} is only flattened and reshaped, so the formula is expressed as follows:

$$X' = \left\{ Proj\left(X_i^{\tau} \right) \right\}_{i \in S} + \left\{ f\left(X_{N-1}^{\tau} \right) \right\} \tag{10}$$

$$Proj\left(X_i^{\tau} \right) = \zeta \left(Linear\left(f\left(X_i^{\tau} \right) \right) \right) \tag{11}$$

where X' represents the set of features after the feature projection of each layer. X_N^{τ} represents the output feature of the fourth stage, f represents the flattening, *Proj* represents the projection layer, *Linear* represents the Linear projection based on *Linear-Gelu-Linear*, ζ represents the matrix operations that reshape, pool, and flatten.

Fusion layer: We adopt the weighted average strategy to fuse the output layer with the output of the previous layers, and map the shallow features to the deep layer to continue learning. During training, the weights are distributed equally to each output layer, then normalized and updated dynamically, and sum to 1, whose absolute value indicates the importance of each layer for the classification task. Then the eigenvalues corresponding to each layer of the stacked matrix are summed to output the final fused features for classification. Given that the classification network has a total of *N (N=S+1)* stages, including *S* projection layers and an output layer, the formula of multi-level feature fusion method based on weight average strategy is defined as follows:

$$O = F\left(X' \right) = \sum_{i \in S} w_i \cdot X_i' + w_{N-1} \cdot X_{N-1}' \tag{12}$$

where F represents the fusion layer, fusing multi-level projection features, X_i' represents the i-th feature matrix in X', w_i represents the weight assigned to each of the S projection layers, w_N represents the weight assigned to the output layer of the last stage, these weights are initialized as 1/N by parameterization method, and are dynamically updated during the training. Its absolute value represents the importance of each layer feature for the classification task, and O represents the output fusion feature.

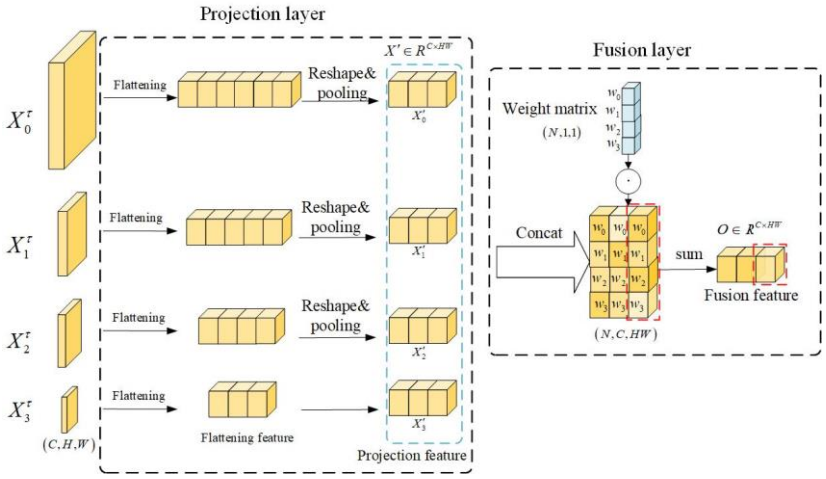

Figure 5: Architecture of multi-level feature fusion based on weight averaging strategy.

3. Experiment and Analysis

This study conducts experiments based on PaddlePaddle framework. The graphics card is an RTX 2080 Ti. The batch size is set to 32 and the epoch is set to 300. In this section, the ablation experiments and comparison experiments with other advanced methods are carried out using the WRF dataset containing winter snow and ice road.

3.1. Dataset and evaluation metrics

WRF dataset is a winter road dataset containing 5061 images proposed by Tongji University in Shanghai, which contains road images under various weather conditions in cold regions in winter. The dataset classifies the road condition into six categories: dry, waterlogged, partially covered, melted snow, snow and muddy road surface.

In this experiment, the number of parameters (Params), computational complexity (Flops), and classification accuracy (Acc) are selected as evaluation indicators.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

where TP represents a true example, FP represents a false positive example, FN represents a false negative example, and TN represents a true negative example.

3.2. Ablation experiments

In this subsection, four sets of ablation experiments are set up to prove the effectiveness of the proposed Dynamic Overlapping Slice Embedding (DOPE), Frequency-based Decomposition Attention (FFA) and Multi-level feature fusion based on Weight Average strategy (WAFF) methods. Table 1 shows the results of the experiment. Experiment 1 represents the experiment based on the original model of PVTv2. In Experiment 2, OPE in Experiment 1 is replaced by DOPE method, and the Acc is improved by 3.26%. In Experiment 3, our proposed FFA method was used to further replace the linear attention method in the original model, and the results show that Acc was improved by 3.65%, which proved that FFA method had better performance than linear attention. In Experiment 4, WAFF method is added based on Experiment 3, the results show that our method can achieve an accuracy of 88.93% using almost the same Params and Flops as the original model, and the classification accuracy is improved by 8.42%. In summary, the analysis proves the effectiveness of the proposed method.

Table 1: Ablation experimental results of the proposed method on the WRF dataset.

Experiment	DOPE	FFA	WAFF	Params (M)	Flops (G)	Acc (%)
1				3.6	0.57	80.51
2	✓			3.6	0.58	83.77
3	✓	✓		3.7	0.58	86.26
4	✓	✓	✓	3.8	0.60	88.93

3.3. Comparison with other mainstream classification methods

In this section, the effectiveness of the proposed method is demonstrated by comparing it with other mainstream road condition classification methods. Table 2 shows the results of the comparative experiments. The classification accuracy of the improved road condition classification method based on hierarchical Transformer proposed in this study is significantly higher than that of other methods,

and the Parmas and Flops are only 3.8M and 0.60G, which are far lower than most other methods, and can be used as a lightweight classification model. Although Parmas and Flops are slightly higher than PVTv2, the Acc is 8.32% higher than that of PVTv2. Compared with the best STViT method, we still surpass its classification accuracy by 5.08% with a lower Parmas and Flops. This fully proves the effectiveness and superior performance of the proposed method.

Table 2: Experimental results comparing the proposed method with other classification methods.

Methods	Params (M)	Flops (G)	Acc (%)
ResNet101	44.7	7.92	81.53
InceptionV3	27.2	5.71	81.57
Swin Transformer	29.0	4.51	80.15
STViT	52	9.86	84.85
PVTv2	3.6	0.57	80.51
Ours	3.8	0.60	88.93

4. Conclusion

Aiming at the problem that the current road condition classification method for autonomous driving cannot accurately identify the winter road condition with low discrimination, we propose an improved winter road surface classification method based on hierarchical Transformer to improve the classification accuracy of winter ice and snow road. Through experiments on the WRF dataset, the classification method proposed in this paper is superior to the existing mainstream classification methods in terms of computational overhead and accuracy, which provides technical support for safe driving of autonomous driving in winter.

Acknowledgement

This work is supported by the Natural Science Foundation of Heilongjiang Province (LH2022E114), "East Pole" Academic Team Project of Jiamusi University (DJXSTD202417) and Horizontal Project of Jiamusi University (JMSUHXXM2024082101).

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[2] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324).

[3] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[5] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2022). Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media, 8(3), 415-424.

[6] Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., ... & Qiao, Y. (2023). Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12581-12600.

[7] Dong, B., Wang, P., & Wang, F. (2023, June). Head-free lightweight semantic segmentation with linear transformer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 516-524).

[8] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., ... & Yan, S. (2022). Metaformer is actually what you need for vision. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10819-10829).