Research on Lightweight Deep Learning Real-Time Image Recognition Methods for Drone Aerial Photography

Shuqing Li, Mingrui Lai, Mengyao Wang*

Guangzhou Institute of Science and Technology, Guangzhou, 510540, China
Corresponding author: 601877145@qq.com

Abstract: *The problem is that there is little time and the resources on board the drone are very limited. The attributes of aerial images are not uniform; that is to say, there is an inherent trade-off between model accuracy and inference speed, there are differences in perspective and considerable variations in scale, and there is a high density of small targets. Based on the problems mentioned above, some lightweight deep learning methods for real-time image recognition on drones will be introduced in this paper. At the level of the backbone network, in order to improve the feature learning ability and reduce computational cost, we have modified the structure of depthwise separable convolutions and introduced feature reuse and dimensionality reduction. A hierarchical feature adaptation fusion strategy is introduced in the detector stage of this paper to expand the range and increase the expressiveness of features, and also to solve the problem of failing to detect small targets. Channel pruning is performed at the time of deployment, and at the same time, fixed-point quantization and operator kernel tuning are carried out to increase both the model compression ratio and the inference speed; thus, this paper proposes a relatively small, high-accuracy and fast-running solution for real-time drone aerial recognition.*

Keywords: *Drone Aerial Photography, Lightweight Deep Learning, Real-time Image Recognition, Multi-scale Feature Fusion, Model Compression.*

Introduction

Many places for environmental observation and intelligent transportation have started using drones to take pictures and recognise them nowadays. However, due to the limitations of power consumption and storage on the onboard platform, and given that it has very high real-time requirements, the conventional high-precision model cannot be used directly. Aerial pictures are also not very easy to recognise; they are distorted due to the bird's-eye view, have a complicated background, and vary greatly in size according to the height of flight. Although the previous light networks have reduced the computational cost, they still have the problem of multi-scale target representation, cannot find small targets, and are not suitable for inference on embedded systems. Research on Lightweight Deep Learning Real-time Image Recognition Methods for Drones in Aerial Photography is Needed. The three directions of this paper are: Design of a Lightweight Backbone Network, Real-time Multi-scale Feature Fusion, and Model Compression with Inference Optimisation; the goal is to develop a high-accuracy, real-time recognition technology system for resource-constrained devices.

1. Design of Lightweight Backbone Networks for Resource-Constrained Platforms

1.1 Optimisation of the Depthwise Separable Convolution Structure for Drone Aerial Scenarios

The Model Size and Inference Time of the Real-time Drone Image Recognition should be relatively small. A small part of a light-weight neural network is depthwise separable convolution; that is, to reduce the number of parameters and computations of standard convolutions, they are split into depthwise convolutions and pointwise convolutions. However, due to the special imaging characteristics of drone aerial images, such as target deformation in a bird's-eye view, interference from complex backgrounds of ground objects, and variations in target scale caused by changes in flight altitude, there are problems with the feature extraction ability of conventional depthwise separable convolutions. To solve the problem above, we need to change the structure of depthwise separable convolution to reduce the coupling effect between spatial features and channel features, keep the computational cost low, and thus improve the feature representation quality of non-rigid targets in aerial images.

The two ways to achieve this are to change the size of the convolution kernel and to add cross-channel information interaction. At the level of the convolution kernel, the above method introduces the concept of deformable convolution to replace the traditional fixed-geometry depthwise convolution kernel; that is to say, the sampling points are shifted adaptively according to the target deformation to better capture changes in the posture of the target from the drone's viewpoint. Therefore, at the level of pointwise convolution, we can add local dependency information to improve the features fused from different channels and keep the computational cost low. For example, one can use group convolution and channel shuffling to exchange information among the different groups of channels. The structure optimisation above can help to solve the problem of large geometric distortion in complex aerial images and also keep the model light[1].

1.2 Feature Reuse and Dimensionality Reduction Mechanisms for Multi-scale Target Representation

Drone aerial images have a very large scale, and the objects in these images are also very diverse; some are large buildings that cover most of the image, and others are distant vehicles that are only a few pixels in size. The first feature pyramid can solve the problem of scale by combining features at different levels, but the resulting output is still too large for the computation, and thus it is not suitable for resource-constrained drones. To increase the feature hierarchy and reduce the computation, feature reuse is employed to connect the shallow positional detail information with the deep semantic abstraction information at all levels. In the Design of Light-Backbone Networks, this can be done in many ways, such as all sorts of dense connections or skip-layer connections, and each layer obtains gradient information from the previous layer to improve the efficiency of gradient propagation and feature utilisation under parameter constraints.

Reduce the volume and count of data to increase the speed of computation. If there are many feature map channels, a reduction module is added to reduce the amount of redundant information and thus decrease the computation cost of the following layers. In particular, it can be done in a bottleneck structure; that is, first reduce the channel dimension with pointwise convolution, then extract spatial features using depthwise convolution, and finally expand the channel dimension again via pointwise convolution. A small structure that first decreases and then increases in size can reduce the amount of calculation and still have many features. A Channel Attention Mechanism can be used to give different weights to the feature maps at various scales, improve the efficiency of information retention in the dimensionality reduction stage, and thus obtain good speed and high accuracy for multi-scale target recognition by the model.

1.3 Lightweight Feature Recalibration Module with Attention Mechanism

Drones are used to take aerial pictures of all sorts of places under different circumstances, and there will be changes in light and other obstructions. Therefore, the model should be able to concentrate on some important parts of the features. Model the distribution of the importance in the feature channel or at different locations to have the attention mechanism focus more closely on relevant information for recognition. In the Design of lightweight networks, the added attention module should have a small number of parameters and low computational complexity so as not to reduce the lightweight nature of the backbone network due to excessive module complexity. Therefore, the Design of the Light-Weight Feature Recalibration Module with an Attention Mechanism should be able to carry out feature calibration efficiently and have a small computational cost.

It can be designed to be light and have both channel attention and spatial attention. In the channel dimension, a cascade of global average pooling and 1D convolution is used to generate channel weight vectors, and fully connected layers have been replaced to reduce the number of parameters and achieve dynamic recalibration of the importance of different feature channels. Depthwise convolution and group normalization are used to form a spatial attention map in the spatial direction, and this map can dynamically change the weight given to the response at different positions of pixels in the feature map. Add the attention outputs of both directions to the original feature map in a residual way to get the final reweighted features. A Light-Attention Module can be added as a plug-and-play module to one of the main stages of the backbone network to strengthen the detection of prominent objects in aerial pictures without increasing the total size of the network[2].

2. Real-time Multi-scale Feature Fusion based on a Single-stage Detector

2.1 Hierarchical Feature Adaptive Fusion Strategy for Drastic Scale Variations in Aerial Images

When a drone is flying and changes its height or takes pictures from different angles, the size of the same object in the pictures will be relatively large or relatively small. Therefore, the detector should give more weight to the information at all the above levels. The first kind of feature pyramid network has a fixed top-down path for feature fusion, and its fusion weights and connections are fixed after training; thus, it cannot be dynamically adjusted according to the actual distribution of target scales in the input image. Therefore, we will construct an adaptive fusion scheme for hierarchical features that can dynamically select or combine feature maps at different levels of the network according to the scale characteristics of the target in the aerial scene to obtain more distinct multi-scale representations.

The first section of the adaptation is a trained module that outputs the fusion weight. Take the feature maps from all the levels as input to the module and compute the weights for the features at each level of the current input using a small network. One is to give different weights to the combination of deep semantic features and shallow detail features via an attention mechanism. If there are many large objects in the aerial scene, increase the weight of deep features to obtain more semantic information adaptively. If there are many small objects in a scene, the model will pay more attention to the location information of shallow features. The data-driven fusion method mentioned above can solve the problem of expression limitation for a fixed topological structure, generate fused features that are in line with the target scale under the complex scale distribution of drone aerial photography, and thus improve the adaptability of the model to large changes in scale.

2.2 Expansion of Receptive Field and Feature Refinement for the Problem of Missing Small Target Detection

Small targets in drone aerial images generally have only a few dozen or a few pixels, and their signal in feature maps is easily lost due to background noise during downsampling; thus, the detector has a high rate of missed detections. The reason there is a problem with not detecting small targets is that the design of the receptive field in the standard convolutional neural network does not cover the target area, and there is also a loss of spatial information due to a reduction in feature map resolution. Therefore, in order to solve this problem, both the receptive field and the features of the detector need to be improved; that is to say, it should be able to detect small objects more accurately in real time.

The aim of extending the Receptive Field is to expand the scope of information that can be gathered for finding small objects by a network unit. In short, the two are in parallel or series, and dilated convolution can be used to exponentially increase the receptive field of the convolution kernel without increasing the number of parameters or computations. Different Dilation Rates are employed to build a multi-branch Dilated Convolution Module, and then both short-range detailed information and long-range contextual information can be obtained in this way to improve the recognition ability of small objects in the environment. The reason for the reduction in spatial resolution after downsampling is that it has been reduced. One is to use upsampling methods, such as transposed convolution or sub-pixel convolution, to increase the size of the deep low-resolution feature map and then add it to the shallow features element-wise. Refinement can be employed to correct the attenuation in the response for small targets in the deep feature map, and thus the spatial localisation and semantic information retention of the detector will be improved[3].

2.3 Computational Redundancy Pruning and Parameter Reconstruction Methods for Cross-Layer Connections

The basis for the bottom of the cross-layer connection in multi-scale detection frameworks based on feature pyramids is the combination of high-level semantics and low-level details. Most of the cross-layer connections are relatively dense 1x1 convolutions to match the number of channels, and repeated fusion of multi-scale features involves many operations of feature map copy and concatenation. The extra calculation will increase the inference time of the drone's small computer. To address the above problem, we need to reduce the computational redundancy in the cross-layer connections and then reconstruct the parameters to reduce the size of the model while keeping the feature fusion ability.

A sparse connectivity graph can be constructed to reduce the amount of computation. Therefore, we know how much the cross-layer connections at all levels of the feature pyramid contribute, which branches have a small impact, and dense connections can be converted into sparse ones. In a particular case, sparse training can be carried out according to scaling factors, L1 regularization is added to limit the scale of the connection coefficients, and thus the coefficients of redundant connections will be reduced to almost zero during training; structural pruning will be achieved. First, obtain the evaluation results of the previous model. Delete the unnecessary connections, reset the parameters of the convolution kernel for the remaining connections, and then carry out fine-tuning training. Factorisation of the standard convolution kernel can be achieved through tensor decomposition to reduce the number of parameters and, at the same time, keep the expressiveness of the original convolution. Pruning and reconstruction can be used together to reduce both the computation cost and the storage requirement of cross-layer connection modules, and the multi-scale fusion ability of the feature pyramid can still be maintained.

3. Model Compression and Fast Inference Optimisation on Embedded Systems

3.1 Compact Network Topology Search Based on Channel Pruning

Given that the embedded system of the drone is small, only shallow deep learning models can be used. To reduce the size of the model, channel pruning can be used to remove some redundant feature channels and their corresponding convolution kernels. Traditionally, one would manually set the weight of pruning for channels and then select the channels with the smallest weights or activation statistics for pruning. Generally speaking, the above methods need to be optimised several times to recover the accuracy after pruning, and they cannot guarantee the optimality of the pruned network topology on the target platform. To solve the problem mentioned above, channel pruning and neural network architecture search have been combined; now it is possible to automatically find a small network structure that is suitable for drone aerial tasks in the compression stage and thus achieve joint optimisation of compression ratio and recognition accuracy[4].

That is, a differentiable pruning method can be used to make the channel selection a continuous and trainable variable. A scaling factor is added to each channel in training, sparse regularization is employed to set the scaling factors of redundant channels to zero, and then these corresponding channels are directly pruned after the end of training. Based on the above, the method will be employed to set pruning ratios at different levels in an adaptive manner. Build a search space of many candidate pruning operations, and then, by means of gradient descent, optimise the pruning ratio at each layer so that the resulting compact network can still recognise aerial images; change the topology according to the distribution of features at different levels. Search-driven pruning can be employed to reduce the number of parameter tuning, and a lightweight network structure will be obtained that is in line with the characteristics of drone aerial data and has good computational efficiency and feature representation ability.

3.2 Forward Inference Acceleration with Fixed-Point Quantization and Activation Function Fusion

Quantisation is a way to reduce the number of bits in floating-point numbers to a small fixed-point number; thus, it is reduced in size, and fixed-point arithmetic units in embedded systems can be employed to increase inference speed. Since drones are generally used to take pictures and videos from above, the Quantisation Method should have good computational efficiency and reasonable accuracy. Fixed-point quantization maps the weights and activation values to integers that are 8 bits or less, and then the convolution operation can be carried out by integer arithmetic logic units; thus, both the computation time and power consumption are reduced considerably. However, the uniformisation will lose some of the details in the aerial photographs. Therefore, the method above should be non-uniform quantization or per-channel quantization based on calibration data to keep the variation of the dynamic range among different channels small and maintain sensitivity to small target features.

To raise the speed of inference, fusion of activation functions is used. In most typical Convolutional Neural Networks, after a convolutional layer, there is usually Batch Normalisation and then a non-linear activation function. Floating-point arithmetic has to visit many places in the feature map for the same operation, so it is slow. Fuse the parameters of batch normalisation with the preceding convolution kernel and merge the activation function and the quantization operation at the quantization stage so that a single operator can perform both linear transformation and non-linear mapping. At the time of embedded deployment, the operator fusion method mentioned above will reduce the number of kernel calls and the volume of data transfer, thus fully utilise the efficiency of the hardware pipeline. ReLU and ReLU6 are commonly used as activation functions in drone systems, and a particular lookup table method can be employed to carry out the non-linear operation in the fixed-point domain and further reduce the forward inference time[5].

3.3 Tuning of Convolution Operator Kernels Under Memory Access Cost Constraints

The inference of deep learning models on embedded systems is limited by the number of arithmetic operations, and to some extent, the number of times memory needs to be accessed is also small. Drones have a small memory bandwidth; therefore, the transfer of feature maps is often the bottleneck for real-time operation. To reduce the amount of data transfer from off-chip memory to on-chip cache and improve data reuse and memory access, one can modify the kernels of the convolution operator. Many models for drone aerial recognition have introduced depthwise separable convolutions, and although they have a relatively small computational volume, the proportion of memory-access overhead is still quite large; therefore, to improve the overall inference speed, optimisation of the memory-access pattern is necessary.

Some tuning methods are to cache the block of input feature maps and reuse the convolution kernel. According to the cache size of the target platform, the input feature map is divided into several sub-blocks so that each sub-block can be fully stored in the cache during computation and repeated loading of data from external memory is avoided. At the same time, changing the order in which the weights of the convolution kernel are stored by channel can also increase the cache hit rate. To meet the high-memory demands of some operations, such as pointwise convolution, loop unrolling and vectorized load instructions can be used to make full use of the parallel nature of a SIMD architecture. Operator fusion can be used to combine a number of consecutive convolution operations into a single one with a single kernel and reduce the writing and reading of intermediate results. In order to carry out algorithm-hardware co-design, the cost of memory access should be reduced, and it needs to be guaranteed that the drone's aerial image recognition system can operate in real time on embedded platforms.

Conclusion

The three contents of this paper are: Construction of a Lightweight Backbone Network, Design of a Multi-scale Feature Fusion Detector, and Optimisation of Inference on Embedded Platforms for Real-time Image Recognition of Drone Aerial Photography. To better address the problem of deformation and scaling of objects in aerial images, we have modified the backbone network in this paper; that is, we have improved the structure of depthwise separable convolutions, added feature reuse and dimensionality reduction modules, and introduced a lightweight attention mechanism. Detectors: To expand the Receptive Field and increase the expressive power of features, a hierarchy of feature adaptation fusion is introduced to solve the problem of missing small targets, and computational redundancy is reduced by means of cross-layer connection pruning and parameter reconstruction. Channel pruning and structure search are employed to construct a small network for automatic model deployment; fixed-point quantization and activation function fusion are used to increase the speed of forward inference, and given the constraint of memory access, operator kernel tuning is carried out to make full use of hardware efficiency.

In the future, attention mechanisms will be added to the automated lightweight Design Method based on Neural Architecture Search, and joint optimisation will be carried out at the compiler and operator levels for specific drone hardware platforms to achieve a good balance of accuracy and speed.

Fund Projects

A Research on Real-Time Image Recognition Algorithms for Unmanned Aerial Vehicles Based on Lightweight Deep Learning, a General Program in Science and Engineering at Guangzhou Institute of Science and Technology (Project No. XJ2025006701)

References

[1] Ge Yongqi, et al. "Research on Alfalfa Inflorescence Recognition Method and Monitoring System

Based on Drone Imagery and Deep Learning." Transactions of the Chinese Society for Agricultural Machinery, vol. 57, no. 05, 2026, pp. 330-341.
[2] Ji Yueqiang. Research on Drone RF Signal Recognition Method Based on Time-Frequency Analysis. 2025. Shijiazhuang Tiedao University, MA thesis.
[3] Wang Guoshuai, et al. "Small Target Recognition Method for Drone Imagery Integrating Transformer Structure and Attention Mechanism." Journal of Nanjing University (Natural Sciences), vol. 61, no. 02, 2025, pp. 214-222.
[4] Fu Yunkai. Research on Target Recognition Methods for Drone Applications. 2023. Shenyang University of Technology, MA thesis.
[5] Bie Tong. Research on Image Recognition and Tracking Method for Urban Low-Altitude Multi-Rotor Drones. 2023. Jiangxi University of Science and Technology, MA thesis.