YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information

Chien-Yao Wang[1,2], I-Hau Yeh[2], and Hong-Yuan Mark Liao[1,2,3]
[1]Institute of Information Science, Academia Sinica, Taiwan
[2]National Taipei University of Technology, Taiwan
[3]Department of Information and Computer Engineering, Chung Yuan Christian University, Taiwan
kinyiu@iis.sinica.edu.tw, ihyeh@emc.com.tw, and liao@iis.sinica.edu.tw

Abstract

Today's deep learning methods focus on how to design the most appropriate objective functions so that the prediction results of the model can be closest to the ground truth. Meanwhile, an appropriate architecture that can facilitate acquisition of enough information for prediction has to be designed. Existing methods ignore a fact that when input data undergoes layer-by-layer feature extraction and spatial transformation, large amount of information will be lost. This paper will delve into the important issues of data loss when data is transmitted through deep networks, namely information bottleneck and reversible functions. We proposed the concept of programmable gradient information (PGI) to cope with the various changes required by deep networks to achieve multiple objectives. PGI can provide complete input information for the target task to calculate objective function, so that reliable gradient information can be obtained to update network weights. In addition, a new lightweight network architecture – Generalized Efficient Layer Aggregation Network (GELAN), based on gradient path planning is designed. GELAN's architecture confirms that PGI has gained superior results on lightweight models. We verified the proposed GELAN and PGI on MS COCO dataset based object detection. The results show that GELAN only uses conventional convolution operators to achieve better parameter utilization than the state-of-the-art methods developed based on depth-wise convolution. PGI can be used for variety of models from lightweight to large. It can be used to obtain complete information, so that train-from-scratch models can achieve better results than state-of-the-art models pre-trained using large datasets, the comparison results are shown in Figure 1. The source codes are at: https://github.com/WongKinYiu/yolov9.

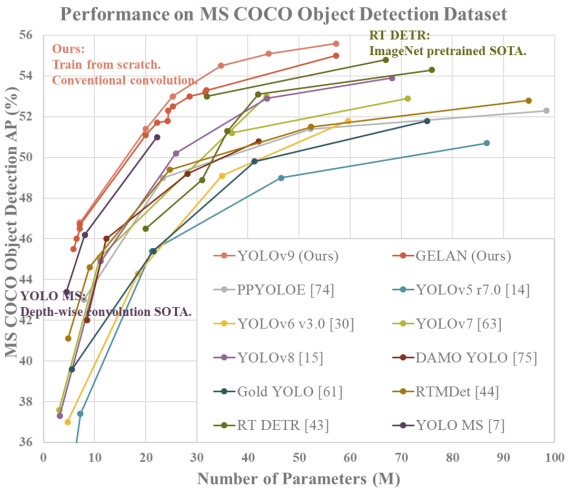
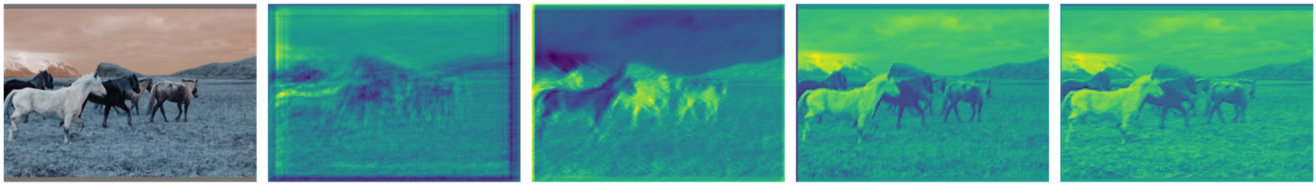


Figure 1. Comparisons of the real-time object detecors on MS COCO dataset. The GELAN and PGI-based object detection method surpassed all previous train-from-scratch methods in terms of object detection performance. In terms of accuracy, the new method outperforms RT DETR [43] pre-trained with a large dataset, and it also outperforms depth-wise convolution-based design YOLO MS [7] in terms of parameters utilization.

1. Introduction

Deep learning-based models have demonstrated far better performance than past artificial intelligence systems in various fields, such as computer vision, language processing, and speech recognition. In recent years, researchers in the field of deep learning have mainly focused on how to develop more powerful system architectures and learning methods, such as CNNs [21–23, 42, 55, 71, 72], Transformers [8, 9, 40, 41, 60, 69, 70], Perceivers [26, 26, 32, 52, 56, 81, 81], and Mambas [17, 38, 80]. In addition, some researchers have tried to develop more general objective functions, such as loss function [5, 45, 46, 50, 77, 78], label assignment [10, 12, 33, 67, 79] and auxiliary supervision [18, 20, 24, 28, 29, 51, 54, 68, 76]. The above studies all try to precisely find the mapping between input and target tasks. However, most past approaches have ignored that input data may have a non-negligible amount of information loss during the feedforward process. This loss of information can lead to biased gradient flows, which are subsequently used to update the model. The above problems can result in deep networks to establish incorrect associations between targets and inputs, causing the trained model to produce incorrect predictions.



(a) Input Image **(b) PlainNet** **(c) ResNet** **(d) CSPNet** **(e) GELAN**

Figure 2. Visualization results of random initial weight output feature maps for different network architectures: (a) input image, (b) PlainNet, (c) ResNet, (d) CSPNet, and (e) proposed GELAN. From the figure, we can see that in different architectures, the information provided to the objective function to calculate the loss is lost to varying degrees, and our architecture can retain the most complete information and provide the most reliable gradient information for calculating the objective function.

In deep networks, the phenomenon of input data losing information during the feedforward process is commonly known as information bottleneck [59], and its schematic diagram is as shown in Figure 2. At present, the main methods that can alleviate this phenomenon are as follows: (1) The use of reversible architectures [3, 16, 19]: this method mainly uses repeated input data and maintains the information of the input data in an explicit way; (2) The use of masked modeling [1, 6, 9, 27, 71, 73]: it mainly uses reconstruction loss and adopts an implicit way to maximize the extracted features and retain the input information; and (3) Introduction of the deep supervision concept [28,51,54,68]: it uses shallow features that have not lost too much important information to pre-establish a mapping from features to targets to ensure that important information can be transferred to deeper layers. However, the above methods have different drawbacks in the training process and inference process. For example, a reversible architecture requires additional layers to combine repeatedly fed input data, which will significantly increase the inference cost. In addition, since the input data layer to the output layer cannot have a too deep path, this limitation will make it difficult to model high-order semantic information during the training process. As for masked modeling, its reconstruction loss sometimes conflicts with the target loss. In addition, most mask mechanisms also produce incorrect associations with data. For the deep supervision mechanism, it will produce error accumulation, and if the shallow supervision loses information during the training process, the subsequent layers will not be able to retrieve the required information. The above phenomenon will be more significant on difficult tasks and small models.

To address the above-mentioned issues, we propose a new concept, which is programmable gradient information (PGI). The concept is to generate reliable gradients through auxiliary reversible branch, so that the deep features can still maintain key characteristics for executing target task. The design of auxiliary reversible branch can avoid the semantic loss that may be caused by a traditional deep supervision process that integrates multi-path features. In other words, we are programming gradient information propagation at different semantic levels, and thereby achieving the best training results. The reversible architecture of PGI is built on auxiliary branch, so there is no additional cost. Since PGI can freely select loss function suitable for the target task, it also overcomes the problems encountered by mask modeling. The proposed PGI mechanism can be applied to deep neural networks of various sizes and is more general than the deep supervision mechanism, which is only suitable for very deep neural networks.

In this paper, we also designed generalized ELAN (GELAN) based on ELAN [65], the design of GELAN simultaneously takes into account the number of parameters, computational complexity, accuracy and inference speed. This design allows users to arbitrarily choose appropriate computational blocks for different inference devices. We combined the proposed PGI and GELAN, and then designed a new generation of YOLO series object detection system, which we call YOLOv9. We used the MS COCO dataset to conduct experiments, and the experimental results verified that our proposed YOLOv9 achieved the top performance in all comparisons.

We summarize the contributions of this paper as follows:

1. We theoretically analyzed the existing deep neural network architecture from the perspective of reversible function, and through this process we successfully explained many phenomena that were difficult to explain in the past. We also designed PGI and auxiliary reversible branch based on this analysis and achieved excellent results.
2. The PGI we designed solves the problem that deep supervision can only be used for extremely deep neural network architectures, and therefore allows new lightweight architectures to be truly applied in daily life.
3. The GELAN we designed only uses conventional convolution to achieve a higher parameter usage than the depth-wise convolution design that based on the most advanced technology, while showing great advantages of being light, fast, and accurate.
4. Combining the proposed PGI and GELAN, the object detection performance of the YOLOv9 on MS COCO dataset greatly surpasses the existing real-time object detectors in all aspects.