Monocular Depth Estimation via Convolutional Neural Network with Attention Module

Lingling Lan¹, Yaping Zhang º and Yuwei Yang²

¹School of Information, Yunnan Normal University, Kunming 650500, Yunnan, China
²Nantong Institute of Technology, Nantong 226000, China
Email: 1624015439@qq.com; zhangyp@ynnu.edu.cn; 20180044@ntit.edu.cn

Abstract. The depth accurately estimated from the RGB image can be used in applications such as 3D reconstruction and scene display. In depth estimation, Convolutional Neural Network (CNN) plays an important role. However, most of the existing CNN-based depth estimation methods often do not take full advantage of the space and channel information in the local receptive field, resulting in lower-resolution depth maps. Based on the extant Encoder-Decoder network architecture, a simple and effective attention module is introduced to learn the discerning feature and the semantic feature in this paper. First, the intermediate feature map is generated by the encoder, and then the attention module is used to generate the attention maps along the two dimensions of the space and channel of the feature map. Finally, the refined output generated by multiplying the attention maps to intermediate feature maps is input to the subsequent neural network for further work. Our works show that the suggest approach has obtained good achievements on NYU Depth v2.

Keywords. Depth estimation; channel attention; spatial attention.

1. Introduction

Currently, Convolutional Neural Network (CNN), in the field of artificial intelligence, is a basic task. CNN does not require manual intervention in feature extraction and image classification that improves the generalization ability of the work. Therefore, scholars at home and abroad have begun to widely use CNN for monocular depth estimation.

Monocular depth prediction is considered as to predict the continuous depth of each pixel from a single color image [1] in many convolutional neural network methods, Eigen et al. first employed CNN for depth estimation [2]. They suggested to use a dual-network architecture to estimate depth. The coarse depth map generated by the coarse-scale network can be used as additional information, and then concatenated with the RGB image as the input of another network. Laina et al. came up with a monocular depth estimation method based on a residual learning full convolutional network, which uses a deeper neural network architecture and does not require post-processing [3]. Alhashim et al. used an encoder-decoder architecture for depth prediction, which can capture more contextual information [4]. Although these jobs, in the image processing field, have achieved good performance, there is still much room for improvement in the quality and resolution of depth maps.

In the human visual system, people are used to capturing high-level semantic information. In the human perception system, people often do not deal with a complete scene at once. On the contrary, in order to get the truth structure, People use local information to observe images. Hu et al. presented a compact module (Squeeze-and-Excitation Networks, SENet) to use the relation between feature
mapping channels [5]. In their “squeeze-excitation” module, the channel-wise attention is calculated by using the global average pool feature. Wen Jing et al. combined SENet and residual network to obtain the depth of color image [6]. Their network architecture shows better performance, but ignores the spatial information of feature mapping. Woo et al. integrated spatial attention and channel attention into one module (Convolutional Block Attention Module, CBAM) and used it for image classification [7]. Wang et al. combined CBAM and residual network to classify images. The experimental results show that using CBAM module has a better effect on image classification [8].

Inspired by Wang et al., this paper combines CBAM and encoder-decoder architecture to perform monocular depth estimation, and uses the attention mechanism to tell the convolutional neural network where and what to pay attention to. It emphasizes useful information and suppresses useless features. Through the test on the NYU Depth v2, the proposed method has improved accuracy that shows the effectiveness of the attention module.

2. Method

2.1. Attention Module

CBAM [7] focuses on channel knowledge and spatial knowledge, and figure 1 shows its network architecture.

![Figure 1. CBAM module [7].](image)

where $I \in \mathbb{R}^{H \times C \times W}$ is an intermediate feature map, the module successively derives an attention map $I_c \in \mathbb{R}^{C \times 1}$ and attention map $I_s \in \mathbb{R}^{1 \times H \times W}$. The attention mechanism can be summarized as equations (1) and (2) [7]:

$$I' = I_c(I) \otimes I$$  
(1)

$$I'' = I_s(I') \otimes I'$$  
(2)

where $C$ is channel of number, $H$ and $W$ are the length and width of the feature map. $\otimes$ denotes element-wise multiplication.

In order to effectively obtain the channel attention, CBAM compresses the spatial dimension of the input feature map. For the aggregation of spatial information, it adopts two operations of global maximum pooling and global average pooling. The channel attention module can be sum up as equation (3) [7]:

$$I_c(I) = \sigma(I_0 \delta(I_0(I_{avg}))+ I_1 \delta(I_0(I_{max})))$$  
(3)

where $I_{max}$ is max-pooling, $I_{avg}$ is max-pooling. $I_0 \in \mathbb{R}^{C \times C}$, $I_1 \in \mathbb{R}^{C \times C}$ are the multi-layer perceptron (MLP) weights. The $r$ is empirically set as 16. $\delta$ and $\sigma$ represent ReLU activation function and sigmoid activate function, respectively.

After getting channel-refined feature $I'$, CBAM first generates two attention maps: $I'_{avg} \in \mathbb{R}^{1 \times H \times W}$ and $I'_{max} \in \mathbb{R}^{1 \times H \times W}$, then concatenates them to become a new feature map. The spatial attention is
summarized as equation (4) [7]:

$$I_s(F) = \sigma(f^{7 \times 7}([I_{avg}, I_{max}])))$$

(4)

where $f^{7 \times 7}$ represents an operation caused by CNN with the filter size of 7×7.

2.2. Overall Network Architecture
The global framework, in figure 2, has two components: Encoder and Decoder. Inspired by Alhashim et al., the encoder is composed of pre-trained DenseNet-169 [9]. CBAM can be inserted into any part of the CNN. According to the experiment, we add three CBAM attention modules to the decoder to take full advantage of the information in the channel and space.

![Figure 2. The overall network architecture.](image)

2.3. Loss Function
We use end-end training procedure. To optimize the entire network, we use equation (5) as the loss function of this article [4]. $\hat{d}$ is predicted depth image and $d$ is ground-truth depth image:

$$L(\hat{d}, d) = L_{\text{grad}}(\hat{d}, d) + L_{\text{sim}}(\hat{d}, d) + \lambda L_{\text{depth}}(\hat{d}, d)$$

(5)

The first term in equation (5) is the $L1$ loss of gradient of the depth images:

$$L_{\text{grad}}(\hat{d}, d) = \frac{1}{n} \sum_{p} g_x(d_p, \hat{d}_p) + g_y(d_p, \hat{d}_p)$$

(6)

where $g_x$ and $g_y$, respectively, calculate the differences in $x$ and $y$ directions for the depth image gradients of $d$ and $\hat{d}$.

Equation (7) represents the loss between $d$ and $\hat{d}$ on the structural similarity (SSIM) [10]:

$$L_{\text{SSIM}}(\hat{d}, d) = \frac{1 - \text{SSIM}(\hat{d}, d)}{2}$$

(7)

Equation (8) describes the difference in the depth values of $d$ and $\hat{d}$. $n$ is the total number of pixels in the depth image:

$$L_{\text{depth}}(\hat{d}, d) = \frac{1}{n} \sum_{p} |\hat{d}_p - d_p|$$

(8)

3. Experimental Result
3.1. Data Set
We use an indoor data set that is NYU Depth v2, which gives RGB images and depth images with a
resolution of 640×480 [4]. In this paper, the job produces depth map that the resolution is 320×240.

3.2. Implementation Details
Our depth prediction network is implemented in TensorFlow and trained on a GeForce GTX 1660Ti GPU for 60 epochs. The encoder of this paper is DenseNet-169 pre-trained on ImageNet. The batch size is set to 2.

3.3. Evaluation
We use six standard indicators to quantitatively evaluate the performance of the trained model. These indicators are defined as:

- **Average relative error (rel)** [2]:
  \[ \frac{1}{n} \sum_{p} \left| \frac{d_p - \hat{d}_p}{d_p} \right| \]

- **Root mean squared error (rms)** [2]:
  \[ \sqrt{\frac{1}{n} \sum_{p} \left( \frac{d_p - \hat{d}_p}{d_p} \right)^2} \]

- **Average (log10) error** [2]:
  \[ \frac{1}{n} \sum_{p} \log_{10} \left( \frac{\hat{d}_p}{d_p} \right) \]

- **Threshold accuracy (δ)** [2]:
  \[ \max \left( \frac{\hat{d}_p}{d_p}, \frac{d_p}{\hat{d}_p} \right) = \delta < thr \text{ for } thr = 1.25, 1.25^2, 1.25^3 \]

where \( d_p \) is a pixel in the depth map \( d \), \( \hat{d}_p \) is a pixel in the predicted depth map \( \hat{d} \), and \( n \) represents the total number of pixels in each depth map. The network in this article is compared with other methods, and the performance is shown in table 1.

**Table 1.** The difference results on NYU Depth v2 data set.

| Method                | δ_1 | δ_2 | δ_3 | rel | rms | log_{10} |
|-----------------------|-----|-----|-----|-----|-----|----------|
| Eigen et al. [2]      | 0.769 | 0.950 | 0.988 | 0.158 | 0.641 | -        |
| Laina et al. [3]      | 0.811 | 0.953 | 0.988 | 0.127 | 0.573 | 0.055    |
| MS-CRF. [1]           | 0.811 | 0.954 | 0.987 | 0.121 | 0.586 | 0.052    |
| Hao et al. [11]       | 0.841 | 0.966 | 0.991 | 0.127 | 0.555 | 0.053    |
| Fu et al. [12]        | 0.828 | 0.965 | 0.992 | 0.115 | 0.509 | 0.051    |
| Alhashim el al. [4]   | 0.837 | 0.967 | 0.992 | 0.131 | 0.570 | 0.055    |
| Ours                  | 0.848 | 0.970 | 0.992 | 0.128 | 0.557 | 0.054    |

Some quantitative results in table 1 refer to the corresponding papers. Due to limited equipment, the result of the method of Alhashim el al. come from the network trained for 60 epochs. After we add the attention module to the network architecture, the accuracy of its performance is improved compared to that of Alhashim et al. [4]. Fu et al. [12] has a large number of epochs, so that its error performance in quantitative evaluation is better than other methods.

The qualitative assessments are shown in figure 3. Among them, the left is the RGB image, the middle is the depth map without the attention module, and the right is the depth map with the attention module. From figure 3, we can find that the network with the attention module can generate better depth maps for images with a smaller depth range, such as the first, third and fourth rows.
3.4. Ablation Studies

This paper uses ablation studies to compare our proposed architecture. Table 2 gives the performance of using different attention modules at the same location in the network of this article. The performance of adding channel attention and spatial attention modules to the network architecture is better than that of only adding channel attention.

| Method       | $\delta_1^{\uparrow}$ | $\delta_2^{\uparrow}$ | $\delta_3^{\uparrow}$ | rel $\downarrow$ | rms $\downarrow$ | log$_{10}$ $\downarrow$ |
|--------------|------------------------|------------------------|------------------------|------------------|------------------|------------------|
| Ours-SENet   | 0.844                  | 0.969                  | 0.993                  | 0.132            | 0.576            | 0.054            |
| Ours-CBAM    | **0.848**              | **0.970**              | **0.991**              | **0.128**        | **0.557**        | **0.054**        |

Figure 3. Visual comparison of estimated depth maps.
Table 3. The different attention modules on NYU Depth v2 dataset.

| Method            | $\delta_1 \uparrow$ | $\delta_2 \uparrow$ | $\delta_3 \uparrow$ | rel $\downarrow$ | rms $\downarrow$ | log$_{10}$ $\downarrow$ |
|-------------------|----------------------|----------------------|----------------------|------------------|------------------|------------------|
| Ours-wo-CBAM      | 0.836                | 0.967                | 0.992                | 0.133            | 0.577            | 0.055            |
| Ours-CBAM1        | 0.849                | 0.971                | 0.992                | 0.127            | 0.558            | 0.053            |
| Ours-CBAM2        | 0.843                | 0.968                | 0.992                | 0.128            | 0.562            | 0.054            |

We add attention modules to different layers of the decoder, and its performance is shown in table 3. In table 3, the first row represents the performance of the decoder without attention module, the second row gives the performance of adding three attention modules to the decoder, and the third row shows the performance of adding five attention modules to the decoder. The results in the table come from the network trained for 40 epochs. Table 3 shows that the performance of appending three attention modules to the network architecture of this paper is better than that of adding five attention modules.

4. Conclusion

We combine an encoder-decoder network with attention module for depth prediction. The attention module is added to the decoder to make the best of the channel information and spatial information and refine intermediate features effectively. With the help of the latest advances in the architecture and the availability of high-performance pre-trained models, our network can achieve better performance on depth map estimation for a single RGB image. Based on this paper’s network architecture, one of the future work is to adopt a more compact encoder for the sake of achieve high-quality depth image prediction on embedded devices. And more clearly identifying the impacts of different encoders, expansion and learning strategies on performance and contribution is also of interest for future work.

Acknowledgments

This work is supported by the Yunnan Ten-thousand Talents Program and Project of Nantong Science and Technology Bureau under grant JC2019108.

References

[1] Xu D, Ricci E, Ouyang W, Wang X and Sebe N 2017 Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5354-5362.
[2] Eigen D, Puhrsch C and Fergus R 2014 Depth map prediction from a single image using a multi-scale deep network arXiv preprint arXiv:1406.2283.
[3] Laina I, Rupprecht C, Belagiannis V, Tombari F and Navab N 2016 Deeper depth prediction with fully convolutional residual networks 2016 Fourth International Conference on 3D Vision (3DV) pp 239-248.
[4] Alhashim I and Wonka P 2018 High quality monocular depth estimation via transfer learning arXiv preprint arXiv:1812.11941.
[5] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 7132-7141.
[6] Wen J and Li Z H 2020 CCDM2020+ 85: Depth estimation method for monocular images based on the Squeeze-and-Control module ResNeXt Journal of Computer Applications 41 (01) 215-219.
[7] Woo S, Park J, Lee J-Y and Kweon I S 2018 Cbam: Convolutional block attention module Proceedings of the European Conference on Computer Vision (ECCV) pp 3-19.
[8] Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X and Tang X 2017 Residual attention network for image classification Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 3156-3164.
[9] Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 4700-4708.

[10] Wang L, Shen X, Zhang J, Wang O, Lin Z, Hsieh C-Y, Kong S and Lu H 2018 Deeplens: Shallow depth of field from a single image *arXiv preprint arXiv:1810.08100*.

[11] Hao Z, Li Y, You S and Lu F 2018 Detail preserving depth estimation from a single image using attention guided networks *2018 International Conference on 3D Vision (3DV)* pp 304-313.

[12] Fu H, Gong M, Wang C, Batmanghelich K and Tao D 2018 Deep ordinal regression network for monocular depth estimation *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 2002-2011.