ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023

Lightweight Human Pose Estimation Based on Self-Attention Mechanism

Youtao Luo [1, a], Xiaoming Gao [2, b]

[1, 2]School of Computer Science and Technology, Southwest University of Science and Technology, China.

[a] iamzhenyu@hotmail.com, [b] carlousjay9011@gmail.com

Abstract. To tackle the issues of numerous parameters, high computational complexity, and extended detection time prevalent in current human pose estimation network models, we have incorporated an hourglass structure to create a lightweight single-path network model, which has fewer parameters and a shorter computation time. To ensure model accuracy, we have implemented a window self-attention mechanism with a reduced parameter count. Additionally, we have redesigned this self-attention module to effectively extract local and global information, thereby enriching the feature information learned by the model. This module merges with the inverted residual network architecture, creating a separate module of WGNet. Finally, WGNet can be flexibly embedded into different stages of the model. Training and validation on COCO and MPII datasets demonstrate that this model reduces the number of parameters by 25%, computational complexity by 41%, and inference time by nearly two times, compared to Hrformer, which also utilizes the windowed self-attention mechanism, at the cost of only 3.5% accuracy.

Keywords: human pose estimation; lightweight; window self-attention; inverted residual network.

1. Introduction

Human pose estimation is a pre-requisite task in many application scenarios such as Human Behavior Analysis [1, 2], human-computer interaction [3], and medical rehabilitation assistance [4] and so on. The main approach is to predict the location of key points of the human body by building neural networks, so proposing models with high recognition accuracy is a hot research topic. Most researchers build deep and complex network models [5-9] to improve the prediction accuracy, however, it also makes the number of parameters larger and the computational speed slower, making it difficult to apply to real scenarios, which is the current problem that needs to solved in the task of human pose estimation.

In order to reduce the number of model parameters, we can cut the depth and width of model directly, but this sacrifices quite a bit of accuracy and therefore we must design the model structure carefully. Some studies [10-12] try to change the model based on complex models and it has some effect. In addition, with Vaswani's [13] self-attentive mechanism dominance on various prediction tasks, more and more researchers try to use it on computer vision tasks. Many studies introduce self-attentive mechanism to the human pose estimation task making the model somewhat lighter because of its strong long-range modelling capability, simple structure and less number of parameters compared to convolutional networks. Li [14] design vision token and keypoint token in Tokenpose, fused them together into self-attentive mechanism block for computation, and combined with convolutional network, which reduce a large number of parameters, but model still suffers from the problem of large computational effort. Yuan [15] proposed a high-resolution network structure in Hrformer and introduce multi-resolution parallel design and local attention, reducing a large number of parameters and computation compared to Tokenpose, but with slow computation speed and long inference time because of its multi-branch parallel design structure.

Above research introduce self-attentive mechanism into human pose estimation and get some lightweight effect, however, it generates square-level space complexity and time complexity when in its computation process, which leading to the problems of slow computation and difficult training. These factors make it still difficult to apply the human posture estimation network to complex vision task scenarios. Therefore we constructed a new attention structure that can efficiently extract

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023

both global and local information. And we apply the new structure to a pose estimation model with lower complexity and shorter inference times than models that also use a self-attentive mechanism. The final inference validation results show that our model still has good detection results.

2. Related Work

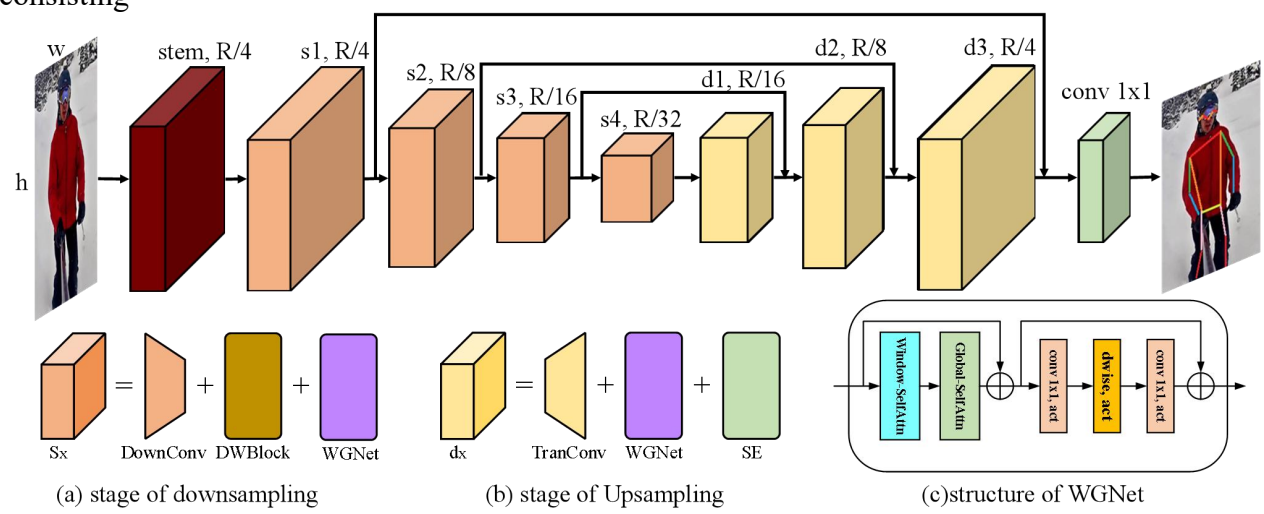
There have been a number of mature studies on Network Lightweight Research. The Mobilenet series [16-18] mainly propose deep separable convolution and inverse residual structure that can reduce a large number of parameters and increase the computational speed, which is the first choice of many lightweight models. Shufflenet [19] group the input features and then channel-mixing wash them to reduce the computational effort while ensuring that the information in each group flow.

Currently, human pose estimation is still mainly done through convolutional networks to predict each key point and many researches introduced transformer[13] into pose estimation and even gradually replace convolutional networks, reducing the number of model parameters compared to convolutional networks, but there is still the problem of low computational efficiency. In addition, some studies aim at lightweight self-attention methods. A representative one is the window-attention of Swin transformer [21], in which the whole feature map is divided into multiple small windows of the same size, and each window completes the calculation of self-attention independently, thus improving the computational efficiency of the model. Hrformer [15] introduce this mechanism into multi-scale fusion model by applying it to semantic segmentation and pose estimation, reducing the number of parameters and computational effort. Methods Hrformer [15], Lite-hrnet [20] are still limited in terms of computational speed because of their multi-branch structure, mainly because the operations of each branch are performed independently and cannot be parallelized. So the task of lightweight human pose estimation task still requires further research.

3. Lightweight Model Structure

3.1 Overall Network Architecture

We comprehensively evaluated the model in terms of parameter count, computational complexity, and inference time, and designed a single-branch lightweight human pose estimation network model.The whole model is shown in figure 1, with a single-way structure of an hourglass and stem layer consisting mainly of depth-separable convolutions with a convolutional kernel step size of 2 and a size of 5. $s_x(x \subseteq [1,2,3,4])$ is the main constituent structure of the model, consisting



(a) stage of downsampling (b) stage of Upsampling (c)structure of WGNet

Fig. 1 Structure of Network

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023
mainly of downsampling in steps of 2, inverted residual networks and WGNet module stacked in sequence. $d_x(x \subseteq [1,2,3])$ recovery the size of feature, using transposed convolution, WGNet and channel attention SE module [22]. Each doubling of the size recovered in this phase is connected with a jump to s_x the corresponding phase, enhancing the return of the gradient during training. The WGNet consists mainly of an improved self-attention mechanism, as shown in Figure 2, stacking the introduced global-attention repeatedly with window-attention, and replacing and replacing the FFN in transformer[13] with an inverted residual network. In the next section, we will analyze in detail the working principle of the self-attention mechanism in WGNet. Finally, we recovery the number of channels to the number of key points by point convolution, and then predict the key point position directly.

3.2 Computational Volume Analysis of Self-Attention in WGNet

The self-attention in WGNet section is shown in figure 2. Firstly, we divide the input feature map into multiple windows of the same size, with each window calculating the self-attention separately. And then take the patch [23] at the same position in each window and stitch together to form a new window, each of which then does the self-attention calculation.

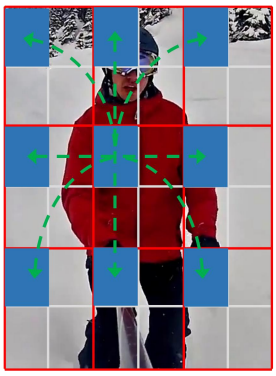


Fig. 2 Principle of the Self-Attention in WGNet

We compare the calculation amount of self-attention in WGNet with the original self-attention (hereinafter referred to as MHSA). Assume that the initial feature height, width and channels are h, w and C. For MHSA, each patch in the feature generates query(q), key(k) and value(v) via the Q, K and V matrices, and let its length remain the same as the input feature depth C, Q, K and V are the parameters to be learned, initialize them to $W_q^{c\times c}$, $W_k^{c\times c}$, $W_v^{c\times c}$. Based on the self-attentive computation process in the transformer[13], plus the computation of generating query(q), key(k) and value(v), the MHSA computation is obtained as follows (here the number of multi-heads is assumed to be 1 and the computation of the softmax function is ignored):

$$\mathrm{Flop}_{\mathrm{MHSA}} = 3\mathrm{hwC}^2 + 2\mathrm{h}^2\mathrm{w}^2\mathrm{C} \tag{1}$$

For self-attention in WGNet, the features are first divided into different windows, let the window size be h' and w', then we get $\frac{\mathrm{hw}}{\mathrm{h}'\mathrm{w}'}$ windows. Then extract patches from the same position in each window form a new window and calculate the amount of computation for each new window according to the calculation steps of MHSA. We let the size of each patch take 4×3, then multiply the number of new windows by the number of calculations per new window, we can obtain the formula for the amount of computation of self-attention in WGNet as follows:

$$\mathrm{Flop}_{\mathrm{WGNet\_attn}} = 6\mathrm{hwC}^2 + 2\mathrm{hh}'\mathrm{ww}'\mathrm{C} + \frac{24\mathrm{h}^2\mathrm{w}^2\mathrm{C}}{\mathrm{h}'\mathrm{w}'} \tag{2}$$

Assume that the input image size is 256×192 and the dimensions are 64×48 after two downsampling layers. The window size needs to be designed to be divisible by the input size. So we

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023

design it to be $4\alpha \times 3\sigma$, which gives the following table comparing self-attention in WGNet with MHSA based calculations.

Table 1. Comparison in Flop of WGNet with MHSA

α	window_size	$Flop_{win\_attn}$	$Flop_{glo\_attn}$	$Flop_{WGNet\_attn}$	$Flop_{HMSA}$
1	(4, 3)	$9.2M+0.7N$	$9.2M+188.7N$	$18M+189.4N$	
2	(8, 6)	$9.2M+2.9N$	$9.2M+47.1N$	$18M+50.1N$	
4	(16, 12)	$9.2M+11.7N$	$9.2M+0.7N$	$18M+23.5N$	$9.2M+18$
8	(32, 24)	$9.2M+47.1N$	$9.2M+2.9N$	$18M+50.1N$	
16	(64, 48)	$9.2M+188.7N$	—	$9.2M+188.7N$	

(attention: M=$1e^3C^2$, N=$1e^5C$)

When α is 1, the window size is (4, 3), which is greater than the calculation amount of HMSA. When α is 2, 4, and 8, the calculation amount is smaller than the latter. When α is 16, the method degenerates into MHSA. When α is 4, the amount of calculation is the smallest, and assume the value of feature dimension C is 32, which can reduce the amount of calculation by 74% compared with the ordinary attention mechanism.

4. Experimental Results and Analysis

4.1 Experimental Environment and Data Set

We use the pytorch framework to build and train the network model on the i7-8700CPU platform. There are 163,961 data in the COCO training set, 149,813 data after cleaning and screening, and 5,000 data in the verification set. The MPII data set has 16 key points marked, 14679 data in the training set, and 2,729 data in the verification set. In addition, we perform various enhancements on images before training, such as random rotation, random cropping, translation transformation, scaling and other operations, which can expand the dataset and also prevent training overfitting.

4.2 Training Results and Analysis

We trained for 200 cycles with an initial learning rate of $1e^{-3}$, and used a cosine learning rate reduction strategy to drop to $1e^{-5}$ at cycle 170, finally obtain 72.1% of mAP in the validation data set. Finally, we compare our method with the current representative method on the COCO and MPII verification sets. The detailed data are shown in Table 2 and Table 3. As can be seen in Table 2, our model has 79% fewer parameters than the HRNet-w32 [9] pure convolutional network,

Table 2. Results Comparison of Different Models on COCO Validation Set

Method	#Params /MB	GFlops	Time/ms	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet-W32	28.5	7.1	36.8	74.4	90.5	81.9	70.8	81.1
TokenPose-S	6.2	11.6	28.7	73.5	89.4	80.3	69.8	80.5
TokenPose-B	13.5	5.7	30.5	74.7	89.8	81.4	71.3	81.4
HRFormer-S	7.8	6.2	70.1	75.6*	90.8*	82.8*	71.7*	82.6*
Ours	5.8↓	3.6↓	24.0↓	72.1	89.6	79.5	68.7	79.6

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023
mainly because the self-attentive mechanism has fewer parameters compared to convolution. Compared with methods using the common self-attentive mechanism [14-15], our improved self-attentive mechanism is less computationally intensive, 36% and 41% lower compared to TokenPose-B [14] and HRFormer-S [15], respectively, and the accuracy is 4% lower compared to HRFormer-S [15], but in terms of inference time spent on a single image, our model inference time of 24.1ms, which is nearly two times shorter than the 74.1ms of HRFormer-S [15]. In Table 3, showing the accuracy of different models for detection at various key points on the MPII data set, our model still has a high accuracy compared to other models and our model has a lower number of parameters. Table 4 shows the ablation experiments based on the self-attentive block in WGNet. The comparison

Table 3. Results Comparison of Different Models on MPII Validation Set

Method	#Params/ MB	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
SimpleBaseline-50	34.0	96.4	95.3	89.0	88.4	88.4	84.0	79.6	88.5
SimpleBaseline-101	53.0	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
HRNet-W32	28.5	96.9	96.0*	90.6	85.8*	88.7	86.6*	82.6	90.1
TokenPose-B	13.5	97.1*	95.8	90.7*	85.8*	89.2*	86.2	82.7*	90.2*
Ours	5.8↓	96.8	95.6	90.4	85.4	88.6	85.6	80.6	89.2

shows that when the self-attentive block in WGNet is completely removed, the accuracy decreases by 3.9% compared to baseline, indicating that the improved self-attentive can effectively extracting feature information.

Table 4. Ablation Experiments

Model	window-SelfAttn	Global-SelfAttn	#Params/MB	mAP/%
Baseline	✓	✓	5.8	72.1
Model-A	✓	×	4.9	70.4
Model-B	×	✓	4.0	69.8
Model-C	×	×	3.2	68.2

4.3 Visual Analysis of Inference

We select complex scene images for detection in the COCO validation set. Figure 3 and Figure 4 show the detection of key points being occluded and multi-person dense, respectively, comparing with the detection results of Hrnet [9], our model detection results are basically consistent with them, indicating that our network still has high detection accuracy.

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023

(a)dection of Hrnet (b)dection of ours model

Fig 3. Part occlusion detection.

(a)dection of Hrnet (b)dection of ours model

Fig 4. Intensive multi-person detection

5. Conclusion

To address the problems of large number of parameters, computational complexity and long detection time when mainstream human pose estimation networks perform inference, we improve the self-attention module and construct a lightweight model that performs well in the computing process. After training on the COCO2017 dataset for single-image inference, achieve a real-time effect of 24ms, and in practical tests, it can accurately identify, the locations of various key points of the human body in practical test. Further research is needed on how to reduce the number of parameters, computational complexity, and subsequent engineering optimisation for deployment to mobile, while ensuring accuracy.

References

[1] Lv Xinyu et al. Analysis of Gait Characteristics of Patients with Knee Arthritis Based on Human Posture Estimation.[J].BioMed research international, 2022,2022:7020804-7020804.

[2] Amir Nadeem and Ahmad Jalal and Kibum Kim. Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model[J]. Multimedia Tools and Applications, 2021,:1-34.

[3] Liu Junfa. 3d human pose estimation and action recognition for human-robot interaction. Master's thesis, School of Electromechanical Engineering Guangdong University of Technology, 2021.

[4] Huu, Phat Nguyen, Ngoc Nguyen Thi, and Thien Pham Ngoc. "Proposing Posture Recognition System Combining MobilenetV2 and LSTM for Medical Surveillance." IEEE Access 10 (2021): 1839-1849.

[5] Toshev A, Szegedy C.Deeppose:Human pose estimation via deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.2014: 1653-1660.

[6] Tompson J J, Jain A, LeCun Y, et al.Joint training of a convolutional network and a graphical model for human pose estimation[J]. Advances in neural information processing systems, 2014, 27.

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.253.2023
[7] Newell A, Yang K, Deng J.Stacked hourglass networks for human pose estimation[C]//European conference on computer vision. Springer, Cham, 2016:483-499.

[8] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 466-481.

[9] Sun K, Xiao B, Liu D, et al. Deep high resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

[10] Yang Senquan and Wen Jiajun and Fan Junjun. Ghost shuffle lightweight pose network with effective feature representation and learning for human pose estimation[J]. IET Computer Vision, 2022, 16(6):525-540.

[11] Xu Dingning et al. LDNet: Lightweight dynamic convolution network for human pose estimation[J]. Advanced Engineering Informatics, 2022, 54.

[12] Li Yanping et al. Human pose estimation based on lightweight basicblock[J]. Machine Vision and Applications, 2022, 34(1).

[13] Vaswani A, Shazeer N, Parmar N, et al.Attention is all you need[J].Advances in neural information processing systems, 2017, 30.

[14] Li Y, Zhang S, Wang Z, et al. Tokenpose: Learning keypoint tokens for human pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11313-11322.

[15] Yuan Y, Fu R, Huang L, et al. Hrformer:High-resolution vision transformerfor dense predict[J]. Advances in Neural Information Processing Systems, 2021, 34:7281-7293.

[16] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[17] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[18] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceed-ings of the IEEE/CVFinternational co-nference on computer vision. 2019: 1314-1324.

[19] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.

[20] Yu C, Xiao B, Gao C, et al. Lite-hrnet: A lightweight high-resolution network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:10440-10450.

[21] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.