P^2 Net : Augmented Parallel-Pyramid Net forAttention Guided Pose Estimation

Luanxuan Hou, Jie Cao, Yuan Zhao, Haifeng Shen, Jian Tang, Ran He*
Center for Research on Intelligent Perception and Computing, Chinese Academy of Sciences
National Laboratory of Pattern Recognition, Chinese Academy of Sciences
University of Chinese Academy of Sciences
AI Labs, Didi Chuxing

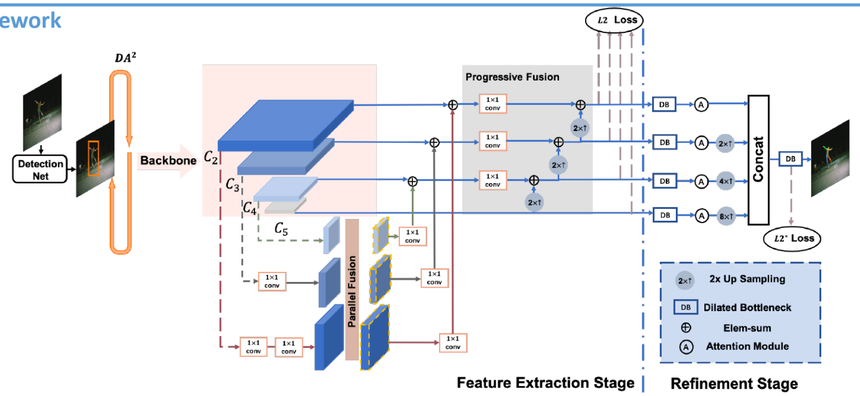
Introduction

- Task: The target of human pose estimation is to determine the body parts and joint locations of persons in the image. Angular changes, motion blur and occlusion in the natural scenes make this task challenging, while some joints are more difficult to be detected than others.
- Challenge: In the unconstrained conditions, the visibility of keypoints is greatly affected by wearing, posture, viewing angle, background etc. The large pose variations further increase the difficulty in detection. Therefore, data augmentation plays an important role in pose estimation. Through data augmentation, we can diversify the data distributions to provide more samples in various scenarios. Existing augmentation strategies are usually hand crafted or based on tuning techniques

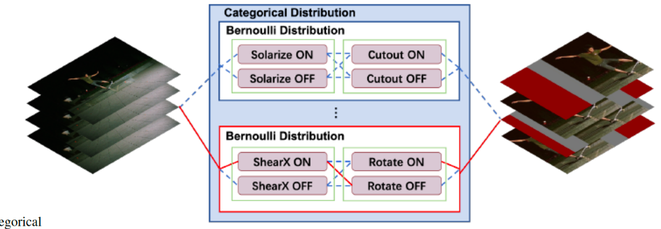
Contributions:

- To the best of our knowledge, we are the first work doing adaptive data augmentation for human pose estimation problem. The data augmentaion policy is learnt to suit specific tasks and training data. Therefore the learnt model is more robust in testing. We formulate the problem of searching data augmentaion policy in a differentiable form, so that the optimal policy setting can be easily updated by back propagation during training.
- We innovate two fusion structures, i.e. Parallel Fusion and Progressive Fusion, to process pyramid features from backbone network. Both fusion structures leverage the advantages of spatial information affluence at high resolution and semantic comprehension at low resolution effectively.
- We propose a refinement stage for the pyramid features to further boost the accuracy of our network. By introducing dilated bottleneck and attention module, we increase the receptive field for the features with limited complexity and tune the importance to different feature channels. Our ablation study shows this refinement stage has a obvious contribution to the overall performance.
- We conducted extensive experiments on the mainstream datasets and achieved the state-of-the-art single model performance. Remarkably, we got 77.3 AP results in the MSCOCO test-dev dataset with smaller Params and GFLOPs.

Framework



Auto Data Augmentation



1) Description of the sampled path: Let C be a Categorical variable with sub-policy probabilities $\pi_1,\pi_2,...\pi_K$.

$$C = one\_hot(\arg_i max(\pi_i)), \quad (1)$$

When training, we want to sample a sub-policy from the fixed *Categorical Distribution* $\pi_1,\pi_2,...\pi_K$ with some randomness. A simple and efficient way to do so is by applying the Gumbel-Max trick [25]

$$C = one\_hot(\arg_i max(\log \pi_i + g_i)), \quad (2)$$

where $g_i = -\log(-\log(u_i))$ are i.i.d samples drawn from $Gumbel(0,1)$. Nevertheless, the description of C in Eq. 2 is discrete. We can't update C using back propagation. Thus we use a generalized form of softmax as continuous and differentiable approximation to $argmax$.

$$c_k = \frac{\exp((\log \pi_k + g_k)/\tau)}{\sum_{k'=0}^{K} \exp((\log \pi_{k'} + g_{k'})/\tau_1)} \quad (3)$$

τ_1 is a temperature parameter. The output of the augmentation is then formulated as:

$$\hat{x} = \sum_k c_k S_k(x) \quad (4)$$

2) Description of the sampled operation: Each operation can be turned on or off inside a sub-policy, which resembles a Bernoulli variable, i.e. b in Eq. 5. Here we ignore the index of operation O, operation probability p and operation magnitude m for simplicity of explanation.

$$\overline{O}(x,p,m) = bO(x,m) + (1-b)x. \quad (5)$$

where $P(b=1) = p$.

Similar to sampling the *Categorical Distribution*, in training time, we want to generate random samples of b from the *Bernoulli Distribution*. After applying the Gumbel-Max trick [26], we have the sample of b as

$$b = H(\log \frac{p}{1-p} + \log \frac{u}{1-u}) \quad (6)$$

where $u \sim Uniform(0,1)$ and H is the unit step function. To make b continuous and differentiable, we relax Eq. 6 using a generalized form of *Sigmoid* function with temperature parameter τ_2

$$b = \sigma((\log \frac{p}{1-p} + \log \frac{u}{1-u})/\tau_2). \quad (7)$$

Quantitative Results

TABLE I
COMPARISONS ON COCO *test-dev*.#PARAMS AND FLOPS ARE CALCULATED FOR THE POSE ESTIMATION NETWORK.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
Mask-RCNN [2]	ResNet-50-FPN	-	-	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI [29]	ResNet-101	353×257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
CPN [6]	ResNet-Inception	384×288	-	-	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [5]	PyraNet	320×256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	-
CFN [30]	-	-	-	-	72.6	86.1	69.7	78.3	64.1	-
CPN [6] (ensemble)	ResNet-Inception	384×288	-	-	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [4]	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W32 [7]	HRNet-W32	384×288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet-W48 [7]	HRNet-W48	384×288	63.6M	32.9	75.5	92.5	83.4	71.9	81.5	80.5
Ours	**ResNet101**	**384×288**	**42.5M**	**26.3**	**77.3**	**93.1**	**84.7**	**73.6**	**83.4**	**82.5**

TABLE II
PERFORMANCE COMPARISONS ON THE MPII TEST SET (PCKH@0.5).

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Stack Hourglass [18].	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Sun et al [31].	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Chu et al [32].	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al [33].	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Yang et al [34].	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al [35].	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang et al [36].	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
SimpleBaseline [4]	98.8	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet-W32 [7]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Ours	**98.8**	**97.0**	**93.9**	**89.9**	**92.1**	**92.0**	**86.4**	**92.9**

Visualization Results