ByteDance
字节跳动

TOUTIAO AI LAB
今日头条人工智能实验室

Team ByteDance-SEU-Baseline

Single-person Human Pose Estimation Track of CVPR'18-LIP Challenge

Speaker: Zhenqi Xu

Kai Su[1,2], Dongdong Yu[2], Zhenqi Xu[2],
Xin Geng[1], Changhu Wang[2]
[1]Southeast University, [2]ByteDance AI Lab

Outline

- Datasets Analysis
- Method Overview
- Single Model Results
- Method Details
- Result Analysis
- Summary
- Future work

Datasets Analysis

Dataset	Number of images	Keypoints
LIP	training 30462 (29866 images is valid), validation 1w, testing 1w * All images are cropped from COCO dataset * The annotation is the same as MPII dataset * The image is already cropped, therefore no person detection is needed.	16
COCO	training 14w+, validation 5k	17
MPII	28881 valid images for training	16

PCKh is used as evaluation measure.

Method Overview

Popular Human Pose Estimation Methods

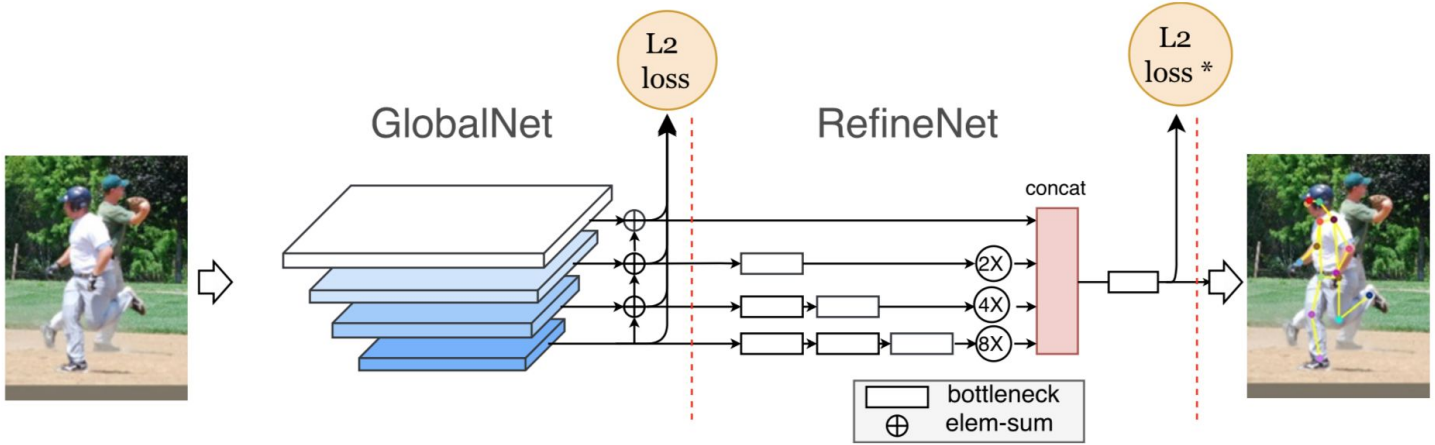
- Stacked Hourglass Networks [1]
- Cascaded Pyramid Networks [2]

[1]*Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European Conference on Computer Vision. Springer, Cham, 2016.*

[2]*Chen, Yilun, et al. "Cascaded Pyramid Network for Multi-Person Pose Estimation." arXiv preprint arXiv:1711.07319 (2017).*

Method Overview

Cascaded Pyramid Networks (CPN)



We adopted CPN, as it performs much better than Hourglass [2].

[2]*Chen, Yilun, et al. "Cascaded Pyramid Network for Multi-Person Pose Estimation." arXiv preprint arXiv:1711.07319 (2017).*

Single Model Results

Team	PCKh on LIP test set
Pyramid Stream Network (Multi-Model) 2nd in the CVPR'17-LIP challenge	82.1
NTHU-Pose 1st in the CVPR'17-LIP challenge	**87.4**
CPN(Resnet-101) trained on LIP trainset	87.0 * batch size is only set to 16, more batch size will perform better.

Method Details

- COCO and MPII Datasets pretraining
- Batch Size is critical
- Ensemble models trained with different backbones

COCO and MPII pretraining is critical

Team	PCKh on LIP test set
Pyramid Stream Network (Multi-Model) 2nd in the CVPR'17-LIP challenge	82.1
NTHU-Pose 1st in the CVPR'17-LIP challenge	87.4
CPN(Resnet-101) trained on LIP trainset	87.0 * batch size is only set to 16, more batch size will perform better.
CPN(Resnet-101) pretrained on COCO and MPII, finetuned on LIP	**89.0** * batch size is only set to 16, more batch size will perform better.

Batch Size is critical

Team	Pre-train	Batch Size	PCKh on LIP test set
NTHU-Pose 1st in the CVPR'17-LIP challenge	-	-	87.4
CPN(Resnet-101)	N	16	87.0
CPN(Resnet-101)	Y	16	89.0
CPN(Resnet-101)	Y	24	**89.8**

Batch Size is critical

Team	Pre-train	Batch Size	PCKh on LIP test set
NTHU-Pose 1st in the CVPR'17-LIP challenge	-	-	87.4
CPN(Resnet-50)	Y	20	89.4
CPN(Resnet-50)	Y	24	89.5
CPN(Resnet-50)	Y	32	**89.6**

Howerver, the performance becomes saturated when increasing batch size.

Ensemble models trained with different backbones

Team	Pre-train	Batch Size	PCKh on LIP test set
NTHU-Pose 1st in the CVPR'17-LIP challenge	-	-	87.4
CPN(Resnet-50)	Y	32	89.6
CPN(Resnet-101)	Y	24	89.8
ensemble CPN(Resnet-50) & CPN(Resnet-101)	Y	32 & 24	**90.2**

Other Details

- Training Augmentation
  - Random scale
  - Flip
  - Random rotation
- Testing Augmentation
  - Flip
  - 40, -40, 20, -20 rotation
- Usually use 4 Tesla-V100 GPUs for training.

Result Analysis

- Details of our submission

Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	UBody	Total
95.800	94.400	91.700	89.600	80.200	89.500	89.200	93.000	90.200

*Hip is much more difficult to be located than other joints.

Summary

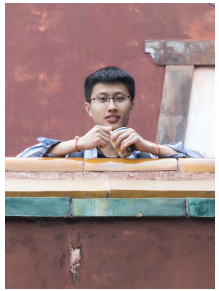
- CPN shows great performance for single pose estimation task.
- Pretraining on the similar datasets is critical.
- Batch size should be large enough.
- Ensemble is critical for higher performance.
- Due to “difficult” joints, more robust architectures are needed.

Future work

When we start trials, there are only about 10 days left, many works are left to do.

- Sync BN layer
- Explore more robust network architectures

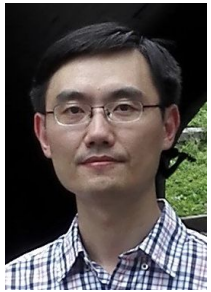
Our team



Kai Su

Dongdong Yu

Zhenqi Xu



Xin Geng

Changhu Wang

Thanks & Questions