Hard Sample Mining Loss for Human Pose Estimation

Tao Hu
Peking University
Email: taohu@pku.edu.cn

***Abstract*—Pose estimation is a typical dense-pixel prediction task. In this paper,We design a novel Hard-Mining loss that specially concentrates on the hard example that is caused by occlusion or complex background, meanwhile gently ignores those easily-discriminated pixel. our contributions are two-folds: 1) We propose a novel loss function named Hard Sample Mining Loss in single-person pose estimation. 2). We give some far-sighted summary on the current single-person pose estimation task.**

I. Introduction

Human Pose Estimation is a fundamental task in current computer vision field. the body parts or limbs with heavy occlusions and extremely cluttered background, however, make this task very challenge. In early years' literature, people often tackle the problem by graph models [1], [2] and random field inference [3], [4] with handcrafted image features. recently, as the the Deep Convolution Neural Network(DCNN) emerged, the performance of a lot of basic computer vision tasks including image classification [5], image segmentation [6], object detection [7] etc have drastically been improved. especially, in image segmentation, [6] proposed a fundamental Fully Convolutional Network(FCN) which radically boosted up the advance of the dense-pixel prediction task such as image segmentation [8], optical flow estimation [9], human pose estimation [10].

A novel architecture named *Stacked Hourglass Network* [10] was proposed to solve the human pose estimation problem. Features are processed across all scales and consolidated to best capture the various spatial relationships associated with the body, the crafty architecture enables repeated bottom-up, top-down inference between different scales for large receptive field. However, Due to person occlusion or some cluttered environment, the localization of joints may be biased or even set to another person. Motivated by the success of the Focal loss [11] which solved the extreme foreground-background class imbalance during training. We design a novel Hard-Mining loss that specially concentrates on the hard example that caused by occlusion or complex background, meanwhile gently ignore those easily-discriminated pixel.

II. Related Work

A. 2D Human Pose Estimation

human pose estimation can be divided into single-person human pose estimation and multi-person human pose estimation. we mainly talk about the single-person case here which is, to some extent, easier than multi-person case. graph structures are often adopted to simulate the spacial relationships between human body joints based on hand crafted features such as SIFT [12], HOG [13]. With the rise of DCNN technology, some models [10] [14] greatly accelerated the performance of human pose estimation. DeepPose [15] is the first architecture that incorporated DCNN in this framework. *Stacked Hourglass Network* [10] is proposed to mix multi-context feature together, and recursively refine the joints predictions via multi-stage intermediate supervision.

B. Hard Negative Mining

Hard negative Mining is very common in diverse computer vision task such object detection [16] [17] [11], image segmentation [18] and image classification [19]. most of the reason of hard negative mining is class imbalance. we have to realize that the gradient is precious, so that we don't hope the gradient finally contributed to the obviously distinguishable sample. On the contrary, those hard negative samples should be taken seriously. the typical solution to proceed hard negative mining are manually class balance [20] [21] and designing a suitable loss that pushes the network to concentrate on the hard samples [11] [21]. Our proposed Hard-Mining Loss can exponentially weaken the influence of the obviously distinguishable sample, meanwhile, augment the effect of the hard samples.

III. The Proposed Method

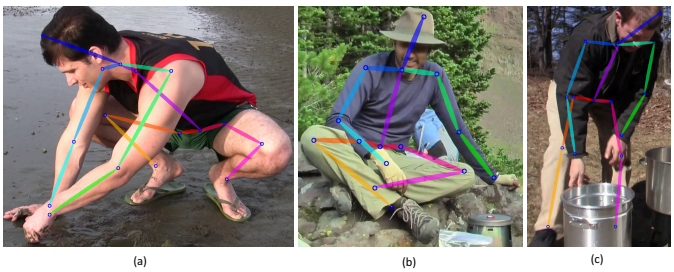

Fig. 1. Result on the MPII dataset. This figure is best viewed in colour. Our method can fit diverse pose about squating (a), seating(b), standing(c). In (c), the left leg can be rightly inferenced even in occlusion.

A. Revisiting Stacked Hourglass Network

Hourglass network structure is demonstrated in Figure 2. it adopts a single pipeline with skip layers to preserve spatial

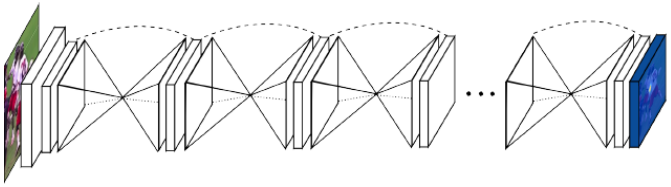

Fig. 2. traditional hourglass architecture. This figure is best viewed in colour.

information at each resolution, therefore can effectively process and consolidate features across scales. At the beginning of the network, two feature map downsampling operations are processed via 7×7 convolution with stride 2, 2×2 max pooling in case that feature maps occupy too much GPU memory in later stage. After that, convolutional and max pooling layers are adopted to transfer the feature to a smaller resolution. after the max pooling operation, the network branches off, the new skip branch is utilized in later upsampling. after reaching the lowest resolution, the network begins the top-down sequence of upsampling and combination of features across scales. Several hourglasses are often stacked together, intermediate supervision is appended after each hourglass. the output of last hourglass is used in inference.

B. Hard Sample Mining Loss

The main form of our loss is a Mean-Squared Error(MSE) loss. The difference lies in that extra Hard Sampling Mining weight is appended on MSE. we name it Hard Sample Mining loss:

$$HSM = (1 - e^{\beta\Delta})\Delta^{\gamma} \tag{1}$$

where Δ denotes the absolute value of the difference between the inferenced feature map and ground truth feature map. and the whole equation is elementwise operation, β, γ are variables that need to be decided by grid search.

$$\Delta = |\sigma(p) - y| \tag{2}$$

Where the p denotes the final feature map output of the whole network, and y is ground truth. The σ is a sigmoid activation function, which is here used to normalize the output range from 0 to 1 for the convenience of β, γ grid search.

C. Training and Inference

We use the final stage feature map to obtain body joints locations. Denote the stage number is S, batch size is N, body joints type number is I. $d_i = (x_i, y_i)$ means the location of joint type i. Firstly we introduce how to construct the ground truth feature map. Directly setting the ground truth feature map a 0-1 binary feature map leads a pool optimization result. Therefore, Gaussian smooth is adopted for every body joint position.

$$F_i \sim N(d_i, \Sigma) \tag{3}$$

where the F_i represents the i-th constructed ground truth label, and the Gaussian variance Σ is set as an identity matrix.

The final optimization loss is as follows:

$$L = \frac{1}{2}\sum_{s=1}^{S}\sum_{n=1}^{N}\sum_{i=1}^{I}||F_i - \hat{F}_i||^2 \tag{4}$$

where $\hat{F}_i = f(x_i, \phi)$, ϕ is the Hourglass network parameter.

During inference, we obtain the predicted body joint location by argmax operation on the final stage's output feature maps, which is formulated as follows:

$$d_i = argmax_p \hat{F}_i(p), i = 1, ..., I \tag{5}$$

where p is predicted position, and $p \in R^2$. $\hat{F}_i(p)$ means activation value at position p of the according feature map.

IV. Experiments

We conduct the experiment on a common human pose estimation benchmarks: The MPII human pose dataset [22], which includes totally 25k images containing nearly 40K people.

A. Implementation Details

Our implementation follows [10]. Training data are augmented by scaling, rotation, flipping and adding color noise. The input image ratio is hold unchanged, conduct padding to make it square ratio. The input image is 256 × 256 cropped by the annotated body position and scale. The Gaussian smooth radius of ground truth is 7. All the models are trained using tensorflow. we use ADAM [23] to optimize the network on one Titan X Pascal GPU with a mini-batch size of 16 for 200 epochs. The learning rate warmup strategy [24] is adopted. The learning rate is initialized as 2.5×10^{-4} and is dropped by half when the objective is stuck in some plateau. We randomly draw 5% of the training data for validation models and always memorizes the best model on this validation set. training a model requires about 4-5 days on a single GPU. testing is conducted on six-scale image pyramids with flipping.

B. Experimental Results

In MPII dataset, the PCKh measure [25] is adopted, where the error tolerance is related to the head size. we use the matching threshold as 50% of the head segment length on the MPII dataset. Some inference results are demonstrated in Figure 1.

as the result shows in Figure 3. the baseline is a 2-stage hourglass structure. others are HSM Loss Hourglass with different β, γ choice. The validation data is chosen from the training data, it contains 480 person. from the Figure, we can deduce the following conclusions.

1) Better Convergence: .in Figure 3 (a)(b) Hourglass with HMS obtain much better accuracy in both training and validation data. furthermore, when $\beta = 1, \gamma = 2$, the evaluation accuracy is largest. which equally represents a exponential weight attached on the standard MSE loss.

2) Higher Accuracy: .in Figure 3 (c)(d), Hourglass with HMS can converge much faster than the baseline. The peak at the beginning of the curve is caused by the warmup [24] strategy.

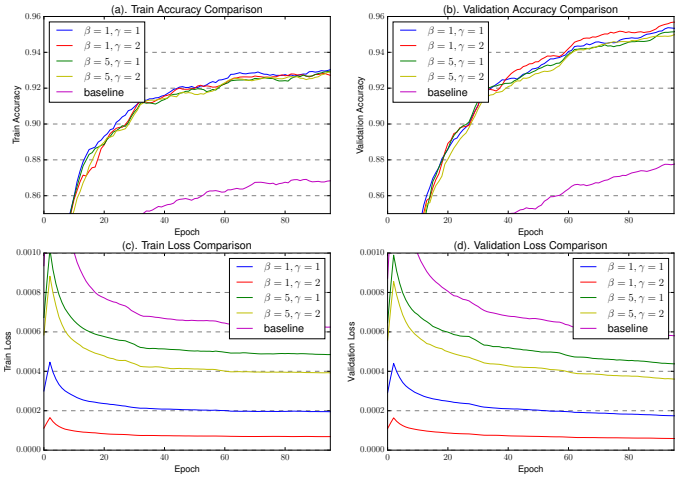

Fig. 3. Train,Validation Accuracy and Loss Comparison. (a). Train Accuracy Comparison. (b). Validation Accuracy Comparison. (c). Train Loss Comparison. (d). Validation Loss Comparison. Note that the curves here are all smoothed by 1D convolution. This figure is best viewed in colour.

V. Result Visualization

A. Common Visualization

Figure 4 is some visualization result between baseline and HSM(best), where HSM(best) means $\beta = 1, \gamma = 2$. we can find the HSM force the neural network to make approximate inference according to the human pose structure constraint. the network will try its best to focus on the hard example that caused by occlusion or complex background, meanwhile gently ignore those easily-discriminated pixel.

Fig. 4. Baseline and HMS(Best) Visualization. This figure is best viewed in colour.

B. Bad Case

Figure 5 is some bad case both in baseline and HMS(best), where HMS(best) means $\beta = 1, \gamma = 2$.

we can find the single person pose estimation exists following problem:

- ankle joint accuracy is obviously too low in comparison to other joints.
- some joints are missing in the image due to the image size. the ideal handling method is marking miss rather than blindly get the max value position in the feature map.
- occlusion is the typical problem in single person pose estimation. when we cannot crop the pure single person when two person are too close to be fully separated. an attention-like mechanism need to be incorporated to force the focus on the most salient person.

Fig. 5. Baseline and HMS(Best) Visualization. This figure is best viewed in colour.

VI. Conclusion

In this work, we propose a novel HSM Loss that not only accelerates the training speed but also boosts the evaluation effect. The implementation detail is provided for future research.

VII. Future Works

Based on the previous summary, We think that it at least exists the following promising possibility of improvement:

- prior information about human pose can be embedded into the framework by Graph Convolution Network [26], Adjoint Matrix etc. thus, the neural network can further inference approximate position even in occlusion.
- ankle-attention module can be adopted to further boost the pool ankle accuracy.
- Some mechanism can be incorporated into the framework, such as setting a feature map activation threshold to avoid "joint missing" problem.

References

[1] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," pp. 1465–1472, 2011.
[2] L. Ladicky, P. H. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," pp. 3578–3585, 2013.
[3] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," pp. 1–8, 2008.
[4] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *European Conference on Computer Vision*. Springer, 2014, pp. 33–47.
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[9] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.

[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," 2017.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[14] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," *arXiv preprint arXiv:1708.01101*, 2017.

[15] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[16] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[17] O. Canévet and F. Fleuret, "Efficient sample mining for object detection," in *Proceedings of the 6th Asian Conference on Machine Learning (ACML)*, no. EPFL-CONF-203847, 2014.

[18] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," *arXiv preprint arXiv:1704.01344*, 2017.

[19] O. Canévet, L. Lefakis, and F. Fleuret, "Sample distillation for object detection and image classification," in *Proceedings of the 6th Asian Conference on Machine Learning (ACML)*, no. EPFL-CONF-203846, 2014.

[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[22] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[25] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[26] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, 2016, pp. 3844–3852.