Paper Reading Report-04

Shreya Chawla
u7195872

Abstract

*This is my reading report for the paper titled: "***Fast Human Pose Estimation***", authored by Feng Zhang (University of Electronic Science and Technology of China) et al, and published in IEEE CVPR 2019.*

All ENGN8501 submissions will be subject to ANU's TurnitIn plagiarism check, against both the original paper, internet resources, as well as all other students' submissions. So please make sure that you must write your own reports, and declare the following originality statement:

I, Shreya Chawla, hereby confirm that I am the sole author of this report and that I have compiled it in my own words.

1. Problem Statement

Human pose estimation refers to the computer vision task to detect semantic key points of a person that is predicting body joints for humans in a given image. This paper is an attempt to efficiently solve the problem with little loss in performance and accuracy.

Pose estimation itself is a complex problem as it involves complicated situations like occlusion and requires high compute power. A scalable algorithm has many practical uses particularly on resource-limited devices such as smart phones and robots. Thus this paper address a very challenging task. It has several applications including human activity estimation in autonomous car, training robots, augmented reality, motion tracking, and human fall detection [3].

One of the current leading technique for an efficient pose estimation is the fast and lightweight pose network [2] which optimizes the OpenPose approach to reach real-time inference on CPU with negliable accuracy drop.

2. Summary of the paper's main contributions

This paper [1] claims that the proposes Fast Pose Distillation (FPD) learning strategy is scalable and faster. The authors state that their network to be more compact than the original redundant Hourglass network. They argue that their method pursues the best model performance given very limited computational budgets using only a small fraction (less than 20%) of cost required by similarly strong alternatives with comparable generalizability. It uses pose knowledge distillation (KD) to transfer the latent information learnt by a pre-trained larger teacher network to a much smaller student network. They claim to first use this technique as FPD which was previously successfully exploited in object image categorisation.

3. Method and Experiment

The FDP consists of the lightweight 4 hourglass pose neural network architecture, with 128 channels per layer, learns features from a larger pre-trained 8 hourglass blocks (256 features each) teacher network using KD technique [1].

The teacher network is built on the original Hourglass and trained on two standard benchmarks datasets - MPII Human Pose and Leeds Sports pose. The supervised model is trained using a set of images labelled with k joints defined in the image space. Objective loss is evaluated using mean squared error (MSE) based loss function that computing error between the predicted confidence map and the actual confidence map.

The lightweight customized Hourglass neural network, trained under the supervision of pre-trained teacher net, learns the latent knowledge transferred at half the original's depth and width. The Objective loss is computed using the loss function $L_{fpd} = \alpha L_{pd} + (1 - \alpha) L_{mse}$. Where, L_{fpd} is the loss against matching the prediction structure of the strong teacher model, L_{pd} is the loss against labelled-ground truth annotations of training samples, α is the balancing weight between the two loss terms estimated by cross validation [1]. During test time, the larger teacher network is discarded as the discriminative information is already transferred to the lightweight neural network. The effectiveness of the model are evaluated on the student network with standard Percentage of Correct Key-points (PCK) metric. To measure the efficiency, FLOPs is used in both training and testing.

Several experiments were carried out by the authors to substantiate their claims. Comparison with a wide range of advanced techniques were performed on the two datasets on perforance metrics like number of parameters and deployment costs along with AUC to provide contrast on accuracy.

They experiment on MSE and cross validation loss demonstrating MSE as a better choice as it is the formulation of conventional supervision loss. Importance of balancing α parameter between the MSE loss and the proposed pose knowledge distillation loss is analysed. Generalization and effect of pose KD are also studied [1].

4. Critical Analysis

4.1. Are the paper's contributions significant?

This paper lays a good foundation towards building compact yet strong human pose deep models. It is the first to use KD for improving scalability of pose estimation models by transferring information of dense joint confidence maps. Prior to this paper, the best alternatives focused on accuracy alone which was achieved at very high costs. The previous attempts at efficiency had dramatic performance drops. They improved their cost-effectiveness by extensively studying redundancy in state-of-art pose CNN architectures thus giving a boost to the deployability for computationally challenged systems [1].

4.2. Validity of the authors' main claims

Extensive comparative evaluation results suggests the method's superiority in model cost effectiveness for LSP dataset. However, for MPII dataset, network by Sekii et al (ECCV 2018) outperforms FDP in deployment cost by 3G. Although compared to the state of art in terms of AUC, it is a massive improvement from 26M to 3M parameters and 63G to 9G deployment cost. The difference in mean accuracy is about 1% in MPII and 4% in LSP datasets. The student netowork achieves 1.0% mean PCKh@0.5 gain similar to the original Hourglass case suggesting its retention of sufficient learning capability. Cost-effectiveness analysis of Hourglass helped formulate a lightweight pose CNN architecture with only 16% (9/55) computational cost but obtaining 98% (90.1/91.9) model performance as compared to the leading designs. This demonstrates their claims.

Their qualitative experiments show that their teacher network can detect problems with ground truth annotations like missing label (ankle joint) and error in labelling the correct position of confidence maps [1]. This supports their claim of generalization at slight cost in performance with drastically lesser number of parameters .

4.3. Limitation and weaknesses

From the qualitative comparisons provided, the method does not generalize well for multi-person pose estimation. This might be due to lack of annotations in the ground truth for all persons present in an image or frame. This can be improved by training on better annotated dataset for multiple people detection at a time like COCO dataset.

Their method also fails for pose estimation with occlusion especially if it is another person. This could be addressed by training a better teacher network or by adding more training examples with occlusion.

4.4. Extension and future work

A better loss function to improve test results could also be investigated. This could also be extended to estimation pose of animals. Future work would entail experiments for a better Hourglass teacher network and better loss function to maximize transfer of latent features. This method could be used for motion recognition tasks in real time. Video datasets could be explored for exploiting spatial information.

4.5. Is the paper stimulating or inspiring?

This paper uses knowledge distillation method for a cost effective pose estimation which was previously used in object detection. The use of several losses and their explanation were interesting. Their finding that many models were using highly redundant hourglass structures resulting in slightly better accuracy with a huge computation cost is instrumental for future development. Further, analysis and comparisons in a wide range experiments gives insights into the choices made by authors.

4.6. Conclusion and personal reflection

To conclude, this paper presents a novel state of the art Fast Pose Distillation learning strategy addressing the understudied and practically significant model cost effectiveness for scalability without accuracy performance compromise. Although with some limitations, this a significant step in efficiency. Extensive comparative evaluation results suggests the method's superiority in efficiency and effectiveness compared to its alternatives supported by the detailed analysis and insights from experiments.

If I were to work on this problem, I would make use a different loss function to ensure better supervision of student network. I would also like to improve multi-person pose estimation.

References

[1] Feng Zhang, Xiatian Zhu and Mao Ye. *Fast Human Pose Estimation*. In Proc. IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR), 2019. 1, 2

[2] H. Ren, W. Wang, K. Zhang, D. Wei, Y. Gao and Y. Sun. *Fast and Lightweight Human Pose Estimation*. In IEEE Access, vol. 9, pp. 49576-49589, 2021. 1

[3] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang and C. Yang. *The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation* In IEEE Access, vol. 8, pp. 133330-133348, 2020. 1