On the Robustness of Human Pose Estimation

[1]Naman Jain† [1]Sahil Shah† [2]Abhishek Kumar [1]Arjun Jain

[1]Department of Computer Science, IIT Bombay, [2]Gobasco AI Labs

{namanjain, sahilshah, ajain}@cse.iitb.ac.in, abhisharayiya@gmail.com

1. Introduction

The past few years have witnessed an exponential growth in the real-world deployment of deep-learning based automation systems. However, alongside their innumerable successes deep- learning systems are extremely prone to adversarial attacks which refer to imperceptible noise that can significantly affect performance! Even in this respect, the study of adversarial attacks on classification systems has seen more activity than regression systems .

Human-pose estimation, referred to as **HPE** for brevity, is one such application that uses a blend of regression and classification approaches to learn the compositionality of human bodies, warranting a separate study. To this end, we present the first comprehensive study of the effects of adversarial attacks on HPE systems and their effectiveness with respect to different design choices like heatmaps vs. direct regression, multi-scale processing, attention and compositional constraints.

2. Experimental Settings

HPE refers to inferring a set of 2D joint-locations or pose from an input image. The first DNN based approach, DeepPose [7], used direct regression to learn the joint locations. Lately heatmaps have become more popular and represent joint-locations with the help of heatmaps, one for each joint, that have Gaussian bumps centered at the corresponding joint location. [5] introduced the stacked hourglass network that feeds the previously predicted heatmaps for further processing with image features sequentially.

In our work, we analyze 5 different architectures - DeepPose [7], Stacked-Hourglass [5], Chained-Predictions [3], Hourglass-Attention [2] and Deeply-Learned-Compositional-Model (or DLCM) [6]. We consider 2 variants of the Stacked-Hourglass consisting of 2 and 8 hourglass stacks, and referred to as 2-SHG and 8-SHG, respectively. We use the MPII dataset to train our models and use relative PCKh [1] (relative degradation w.r.t ground truth) on the validation set as our evaluation metric.

†equal contribution

We consider targeted and untargeted variants of the FGSM and IGSM attacks, and call them **FGSM-U/T** and **IGSM-U/T-N**, where U and T denoted the untargeted and targeted variants, respectively while N stands for the upper limit on the number of iterations in case of iterative attacks. We also consider attacks with the following bound on the l_∞ norm (ϵ): 0.25, 0.5, 1, 2, 4, 8, 16, 32. Besides this, we consider universal adversarial perturbations (referred to as **UAP**) [4] in which we generate an image-agnostic perturbation for a given image distribution.

3. Preliminary Findings

Robustness of Different Models: Based on observations from Fig. 2 and other experiments we find that the relative order of robustness across different architectures is more or less consistent. We observe that heatmap based methods are more robust than direct regression based systems. This is because the direct-regression loss function is also a measure of PCKh after thresholding while heatmap loss produces Gaussian bumps at joint location, which is not as strongly correlated to PCKh. In order to make a fair comparison, we use the same ResNet backbone and use a simple regression loss in one case, and de-conv layers followed by heatmap regression in the other case. We consider variants with and without imagenet pretraining and find that imagenet pretraining increases robustness. We observe that DLCM is most robust, perhaps due to its imposition of human skeleton topology. This encourages further exploration of structure-aware models to counter adversarial examples.

Effect of the Number of Iterations: Fig. 2 plots the relative PCKh for untargeted attacks, for $\epsilon = 8$ with 10 and 100 iterations. We observe that moving from 10 to 100 iterations results in dramatic degradation for all the networks when compared classification problems.

Stacked Hourglass Study: We find increasing the depth of network increases robustness. Next, we study the effect of simultaneous perturbation of outputs of all the stacks of SHG, indicated by suffix ALL, and observe that the attacks become more effective, again evident from 2. This is expected because downstream stacks are supposed to improve upon the predictions of the upstream ones and hence, in-

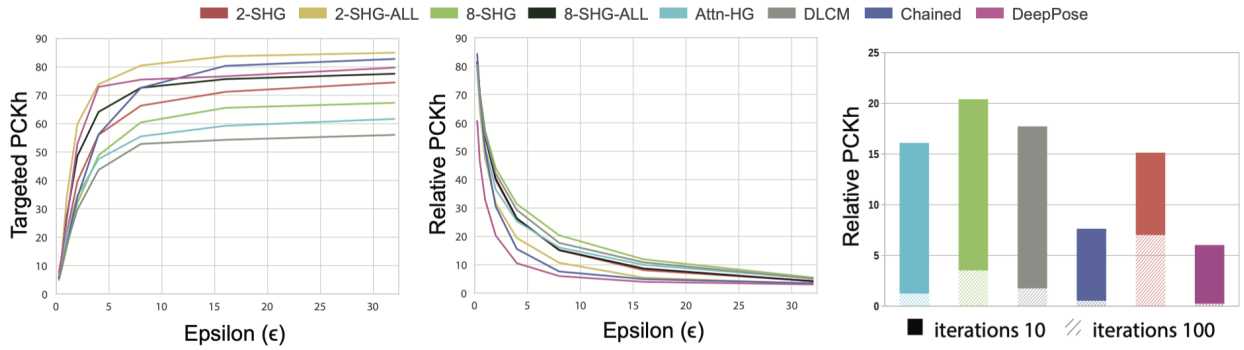

Figure 1. Comparison of different types of attacks on all the models. Left depicts the target PCKh as a function of ϵ for IGSM-T-20. Mid depicts relative PCKh as a function of ϵ IGSM-U-10. Right shows comparision between IGSM-U-10 and IGSM-U-100 attacks

(a) 8-SHG (b) Chained Predictions (c) DLCM (d) DeepPose

Figure 2. Visualization of image-agnostic universal perturbations, with $\epsilon = 8$, for different networks scaled between 0 to 255 for better visualization. Note the hallucinated body-joints, mostly arms and limbs to fool HPE networks. More vis. in supp. mat. Fig 4-6

correct prediction upstream will cascade into errors in the final output. Further, intermediate supervision would provide better gradient flow.

Targeted vs. Untargeted Attacks: Targeted attacks are more difficult than untargeted ones requiring a higher number of iterations. It is because an untargeted attack can simply take large steps in the direction of increasing loss for image I, whereas, the targeted attack requires finding the optimal $I^p : \|I - I^p\|_\infty <= \epsilon$ where the loss is small; a more difficult problem.

Image-Agnostic Adversarial Perturbations: Fig. 2 shows the universal perturbations, scaled between 0 to 255 for better visualization. It is, to the best of our knowledge, the first visualization of such perturbations for HPE, which reveal semantic hallucinations. A closer look reveals that universal perturbations confuse HPE systems by hallucinating body-joints, mostly limbs, throughout the image. Visual inspection of the skeletons predicted on these perturbations reveal similarity with hallucinated joints. Even more interestingly we find that these perturbation are equally effective compared with untargeted attacks, require no computation at runtime and are thus can be used in practice.

Black-Box Attacks: This setting refers to an attack on target network using adversarial perturbations learned from a different network. We observe 30-40% degradation in the target networks performance compared with 85-90% for a white box scenarios. We observe that the generalization is stronger across similar architectures.

Body-Joint Vulnerability Towards Attack: On comparing the robustness of different joints for various architectures, we find that head and neck are the consistently the most robust while hips are the most vulnerable in both targeted and untargeted scenarios. These observations can motivate future work focus on understanding and improving robustness of the more vulnerable joints.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1

[2] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR 2017*. 1

[3] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. 1

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR 2017*. 1

[5] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV 2016*. 1

[6] W. Tang, P. Yu, and Y. Wu. Deeply learned compositional models for human pose estimation. In *ECCV*, 2018. 1

[7] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR 2014*. 1