Kardelen Ceren

CS398 – Project in Computer Science I, 8 ECTS, Spring 2023

EPFL, Visual Intelligence for Transportation (VITA) Laboratory

Professor: Alexandre Alahi, Supervisor: Megh Shukla

Final report

Incremental Learning for Human Pose Estimation

1. Introduction

Human pose estimation is a crucial task in computer vision, involving estimating human body joint positions and orientations from images or videos. It has applications in action recognition, human-computer interaction, surveillance systems, and autonomous vehicles. For example, by accurately predicting the poses of individuals, vehicles can better anticipate their intentions and take appropriate actions, such as adjusting the trajectory or speed for safe navigation.

Traditional pose estimation approaches rely on deep learning models trained on large datasets. These models face challenges when encountering new poses, joints, or environments, as performing well on the new data requires retraining from scratch and substantial computational resources. Moreover, accessing and storing large amounts of old data may pose practical difficulties due to privacy concerns or memory constraints. Fine-tuning the model on the new dataset instead of retraining, however, can lead to catastrophic forgetting, where the model forgets previously learned poses when trained on new ones, limiting its adaptability and flexibility in handling real-world conditions.

To address these challenges, we explore incremental learning techniques for human pose estimation. These techniques enable models to learn from new data while retaining previously acquired knowledge, mitigating the effects of catastrophic forgetting. Our goal is to develop a human pose estimation system capable of incremental learning. As part of our experiments, we investigate regularization with L2 norms and with frequency responses of the convolutional layer kernels.

2. Related Work

In this section, we review existing literature and research related to human pose estimation and incremental learning approaches.

2.1. Stacked Hourglass Model [1]:

The Stacked Hourglass model has been widely adopted in human pose estimation tasks. This model employs a cascaded architecture with repeated bottom-up and top-down layers, allowing for the accurate localization of body joints. While the Stacked Hourglass model is not specifically focused on incremental learning, it serves as a valuable reference for our project's baseline performance. It should be noted that

this model, contrary to common CNN architectures in other computer vision tasks, does not feature fully connected layers after the convolutional layers.

2.2. A Continual Learning Survey: Defying Forgetting in Classification Tasks [2]:

This comprehensive survey on incremental learning investigates three key strategies: replay, regularization, and parameter isolation. Replay strategies involve storing and reusing a portion of the past data samples during training. Regularization techniques introduce constraints or penalty terms that encourage the model to preserve previously learned information while adapting to new data. Finally, parameter isolation methods aim to allocate specific parameters for each task to prevent interference between tasks.

Although focused on classification tasks, the survey provides valuable insights into approaches that can be adapted for incremental learning in the context of human pose estimation, as will be discussed in more detail in the methods section.

2.3. Learning Without Forgetting [3]:

This incremental learning approach for classification tasks introduces a method that uses a regularization term that constrains the parameters of the network based on their importance for previously learned tasks. This approach is not as flexible as regularization methods and requires some architectural change in model training. We instead used this paper for its idea behind using task-specific parameters.

2.4. RBF K-DPP Model Training for Animal Pose Estimation [4]:

Radial Basis Function Kernelized Determinantal Point Process (RBF K-DPP) method focuses on the selection of representative samples from large-scale pose datasets, that the model stores and periodically updates during training. As far as we are aware, this paper is the only one in the field of pose estimation that specifically focuses on incremental learning. However, since this paper assumes some availability of the old data, we decided not to use this method in our experiments.

2.5. Literature Review Conclusion

Based on the literature review conducted, existing incremental learning techniques primarily focus on models with fully connected layers, which poses challenges for popular human pose models that lack such layers. As a result, the transferability and effectiveness of existing incremental learning methodologies in human pose estimation are uncertain. Developing dedicated incremental learning methods for pose estimation may prove beneficial in addressing this limitation.

3. Method Implemented

We modified the model Stacked Hourglass by adding two task heads, with linear layers. After pre-training this base model, we tried three different incremental learning methods: fine-tuning, L2 regularization, and frequency response regularization. Fine-tuning consists of freezing the pre-trained model's parameters and continuing to train the task heads.

As a baseline regularization method, we applied L2 norm regularization to each parameter of the model. L2 regularization is commonly used to prevent overfitting and control the complexity of a model by adding a penalty term to the loss function based on the model's parameters' magnitude. Ours is an adapted

regularization in that loss is computed between the base model's original, unchanged parameters and the new model's trainable parameters. Although this approach was not specifically designed for human pose estimation, it served as a reference for comparison.

Our unique contribution lies in a modified regularization technique. Instead of computing parameter differences, we calculated the differences between the Fourier transforms of the filters in the model's convolutional layers between the original base model and the new model. This approach was tailored to address the specific challenges of human pose estimation.

Fourier transforms allow us to analyze the frequency response of the convolutional filters. By focusing on the changes in the frequency response rather than simply considering the magnitude of individual parameters, we aim to capture the underlying patterns and structures specific to human poses.

4. Experiments

In our experiments, we use PCKh (Percentage of Correct Keypoints with threshold), which is a commonly used metric for evaluating the accuracy of human pose estimation models. It measures the percentage of accurately predicted joints based on a distance threshold.

Following the incremental learning phase, our goal is to ensure that the model's performance on the new dataset is significantly improved without severely compromising its performance on the old dataset.

The above-mentioned base model was pre-trained for 40 epochs, with the MPII dataset, which consists of human pose annotations from real-world images. 10000 images were used as part of the training set.

Then, to simulate the scenario of encountering new data, we employed the LSP dataset with 1000 images, which also provides human pose annotations. We conducted incremental learning experiments using this dataset, each for 30 epochs more, continuing from the pre-trained model. The results can be seen in Table 1. Only left joints' results are reported due to space constraints.

Table 1. Results of different incremental learning experiments, in PCKh@0.53

Model	Training Dataset	Test Dataset	Average	Head	Neck	Left Shoulder	Left Elbow	Left Wrist	Left Hip	Left Knee	Left Ankle
Base Model	MPII	MPII	0.808	0.949	0.952	0.904	0.811	0.725	0.835	0.733	0.677
	MPII	LSP	0.104	0.135	0.147	0.130	0.067	0.055	0.075	0.100	0.074
Fine-tune*	LSP	MPII	0.655	0.902	0.906	0.793	0.630	0.536	0.655	0.542	0.471
	LSP	LSP	0.682	0.867	0.876	0.794	0.582	0.522	0.788	0.628	0.549
L2 Regularization*	LSP	MPII	0.592	0.836	0.846	0.722	0.537	0.462	0.610	0.487	0.479
	LSP	LSP	0.789	0.911	0.916	0.852	0.695	0.637	0.883	0.820	0.765
Fourier Regularization*	LSP	MPII	0.479	0.796	0.805	0.647	0.436	0.361	0.447	0.370	0.300
	LSP	LSP	0.805	0.942	0.944	0.888	0.722	0.653	0.878	0.844	0.712

Models with "*" indicate that they were incrementally trained over the pre-trained base model.

As expected, even though the datasets share the same classes, the base model faces challenges in predicting outcomes for datasets it has never seen, due to differences in data distribution and characteristics. Fine-tuning helps the model adapt to the new dataset without significantly forgetting the base knowledge. However, it may not always achieve optimal learning performance compared to other models. It should be noted that having two heads instead of one helps with fine-tuning as the relevant task head can be affected more.

We observe that both regularization techniques were relatively successful in learning from the new data. While they prevent forgetting to some extent, it can be seen that they both suffered a decrease in the old dataset's accuracy. Regularization with Fourier transforms was worse in terms of retaining previously learned data, despite expectations. Overall, L2 regularization seems to be a better balance between learning and not forgetting.

5. Conclusion

In conclusion, our experiments in incremental learning for human pose estimation have shed light on different techniques and their performance in adapting to new datasets while retaining previous knowledge. We observed that the base model, while capable of predicting the base dataset, faces challenges in generalizing to new datasets. Fine-tuning the model with task-specific heads proved to be a viable approach, as it prevented catastrophic forgetting and allowed for some adaptation to the new dataset, although it did not achieve optimal learning performance for the new datasets.

Regularization techniques, such as L2 regularization, showed potential in mitigating forgetting and preserving base model knowledge. However, it was not without limitations, as some degree of forgetting still occurred. Our novel approach with Fourier transforms of convolutional layer filters exhibited promising results in learning; however, it was not very effective in reducing forgetting.

Overall, our findings highlight the importance of carefully selecting incremental learning techniques based on the specific requirements of the task. It also emphasizes the need for continued research and development to refine these techniques further. In the context of human pose estimation, customizing and combining incremental learning methods can yield better results.

In future work, exploring task heads implemented as convolutional layers instead of linear layers may hold promise for improving human pose estimation models. This approach may enable better capture of spatial dependencies, and lead to more efficient incremental learning approaches.

References

[1] Newell, Alejandro, et al. “Stacked Hourglass Networks for Human Pose Estimation.” ECCV 2016, https://doi.org/10.1007/978-3-319-46484-8_29.

[2] Delange, Matthias, et al. “A Continual Learning Survey: Defying Forgetting in Classification Tasks.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, https://doi.org/10.1109/tpami.2021.3057446.

[3] Li, Zhizhong, et al. “Learning without Forgetting.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8107520.

[4] Nayak, Gaurav Kumar, et al. “Incremental Learning for Animal Pose Estimation Using RBF K-DPP.” BMVC 2021, https://www.bmvc2021-virtualconference.com/assets/papers/0912.pdf.