Deep Learning, without the GPU!

The Effects of Component Size on Model Accuracy, Latency and Memory Usage for mmWave Human-Pose Estimation

Author
Robin Verver

Supervisors
Marco Zúñiga Zamalloa, Nicole Rosi

01 - Background

Human-Pose Estimation describes any technology that can sense the pose of a subject, such as motion capture. Pose estimation with mmWave is done through deep-learning models, however these tend to cost too much computing power, causing high latency or requiring GPUs.

02 - The Model

The model consists of three main components. We don't know which of these contributes most to accuracy, and which might be oversized and thus wasting compute time. In this study we want to find which layers can be reduced without costing accuracy.

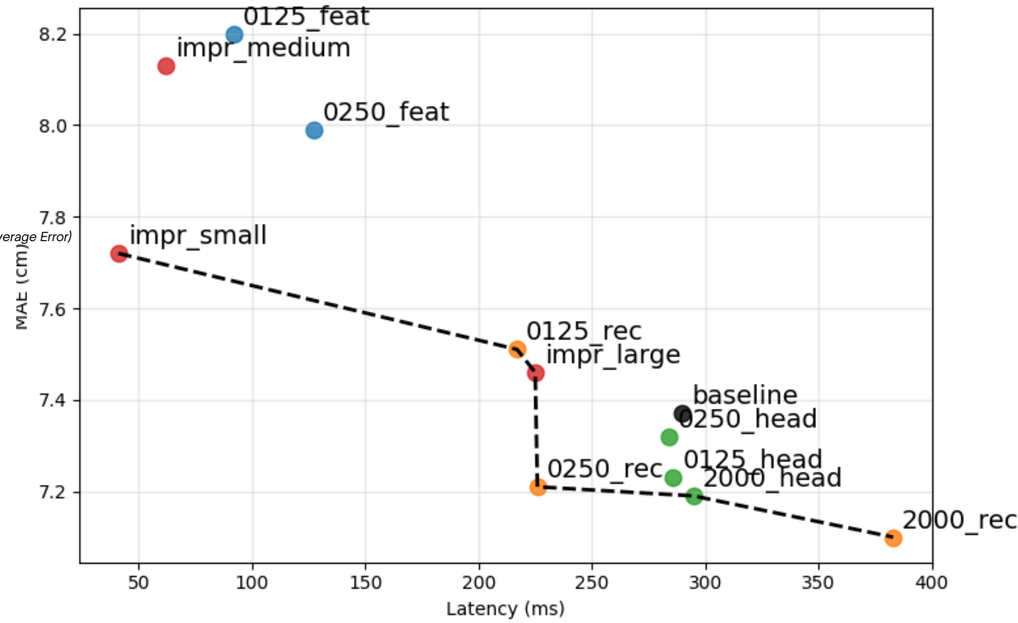
03 - Methodology

In order to find the relative importance of each component on latency and accuracy, we train multiple configurations where we reduce/increase the size of each component, and using these results we suggest 3 new improved models.

Model name	Layer size				
	Feat. Extr.			Rec. Layer	Reg. Head
baseline	32	48	64	64	128
0125_feat	**4**	**6**	**8**	64	128
0250_feat	**8**	**12**	**64**	64	128
2000_feat	**64**	**96**	**128**	64	128
0125_rec	32	48	64	**8**	128
0250_rec	32	48	64	**16**	128
2000_rec	32	48	64	**128**	128
0125_head	32	48	64	64	**16**
0250_head	32	48	64	64	**32**
2000_head	32	48	64	64	**256**
impr_small	**4**	**6**	**8**	**16**	**16**
impr_medium	**8**	**12**	**16**	**16**	**16**
impr_large	**32**	**48**	**64**	**16**	**16**

04 - Results

- Feature extraction seems to impact latency and accuracy the most
- Recurrent layer has no clear relation to accuracy, but costs latency
- Extraction head minimally impacts both latency and accuracy

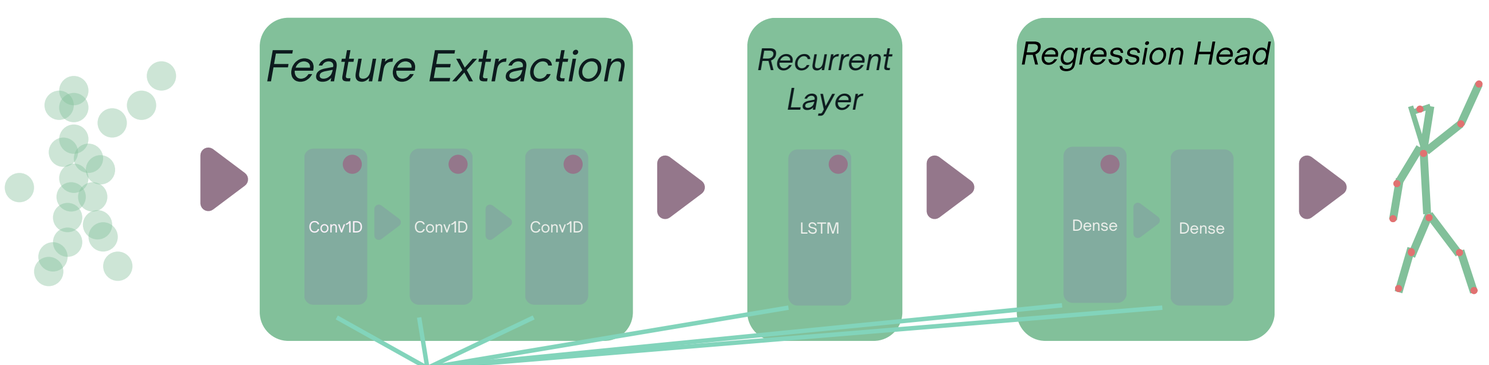


05 - Conclusion & Future Work

We propose two new models: impr_small and 0250_rec. 0250_rec is an overall improvement over the baseline, and impr_small saves significant (85,9%) latency at the cost of some accuracy (4.8%)

The recurrent layer seems to not influence performance as much as expected, and thus future research could investigate if the component is malfunctioning.

Future research might also be interested in the relationship between the sizes of the different layers inside each component, instead of just between the components



Neural network layers, with top-right dot indicating they can potentially be reduced

TU Delft