Radar based Gesture Recognition

Lakshya Kumar Tangri 2018EE10222
Somanshu Singla 2018EE10314

Faculty Supervisor: Prof. Seshan Srirangarajan

19 April 2022

Introduction

- Human gestures: a more natural interface between humans and computers.
- Cameras: Alone or together with other devices raise major privacy concerns

Short-range radars have the ability to detect micro-movements with high precision and accuracy making it a good candidate for realizing gesture-based interfaces

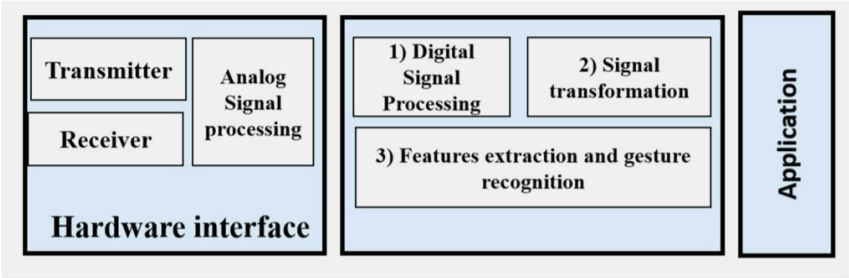
---

Source: Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: ubiquitous gesture sensing with millimeter wave radar. ACM Trans. Graph. 35, 4, Article 142 (July 2016), 19 pages. DOI:https://doi.org/10.1145/2897824.2925953

Project Objective

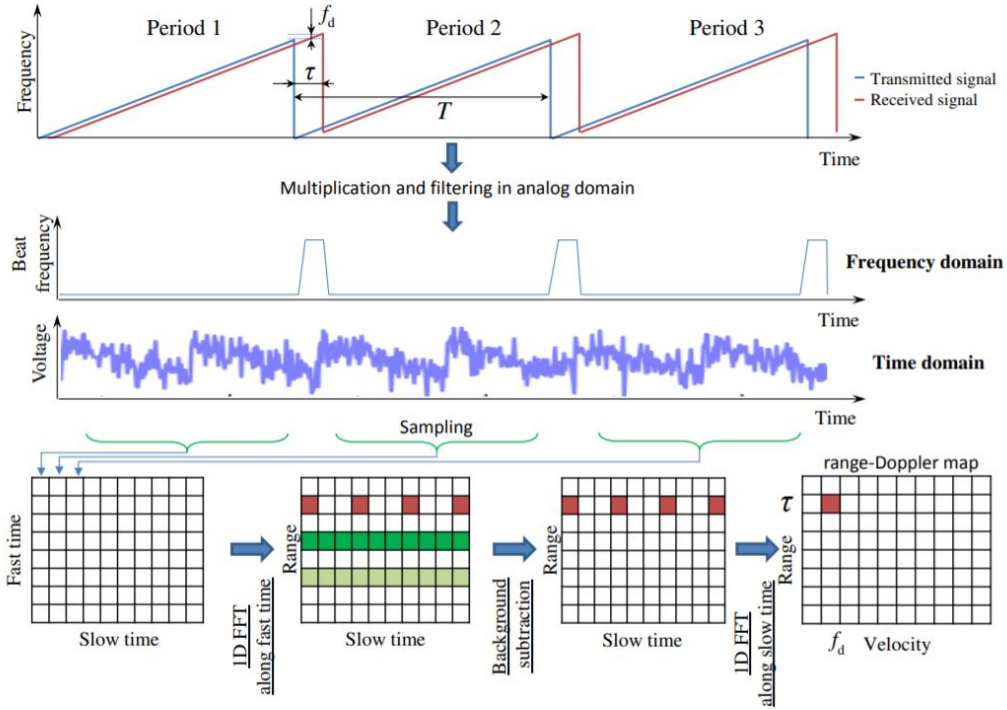
To create a robust, radar based power efficient system which can classify predefined gestures accurately in real time on an edge device

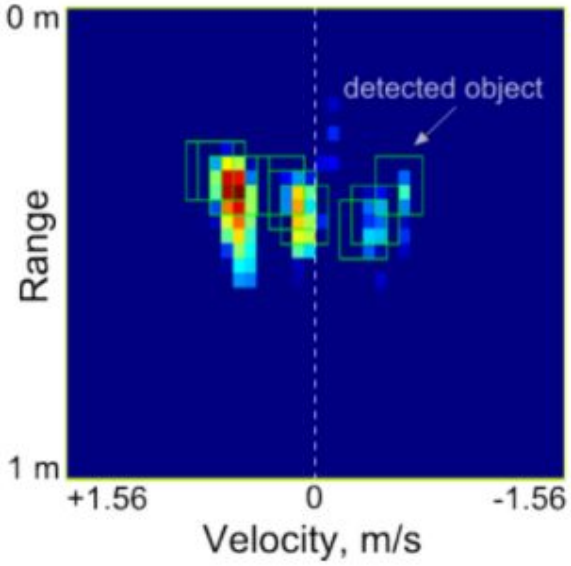
- Robustness: With respect to different users(inter-user) and different gesture instances of a single user(inter-session)
- Accuracy: To be able to correctly classify the input gesture
- Power Efficiency and Real Time: Ability to recognize gesture quickly for a seamless user experience without draining too much battery



Source: Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: ubiquitous gesture sensing with millimeter wave radar. ACM Trans. Graph. 35, 4, Article 142 (July 2016), 19 pages. DOI:https://doi.org/10.1145/2897824.2925953

Range Doppler





Source: P. Molchanov, S. Gupta, K. Kim and K. Pulli, "Short-range FMCW monopulse radar for hand-gesture sensing," 2015 IEEE Radar Conference (RadarCon), 2015, pp. 1491-1496, doi: 10.1109/RADAR.2015.7131232.

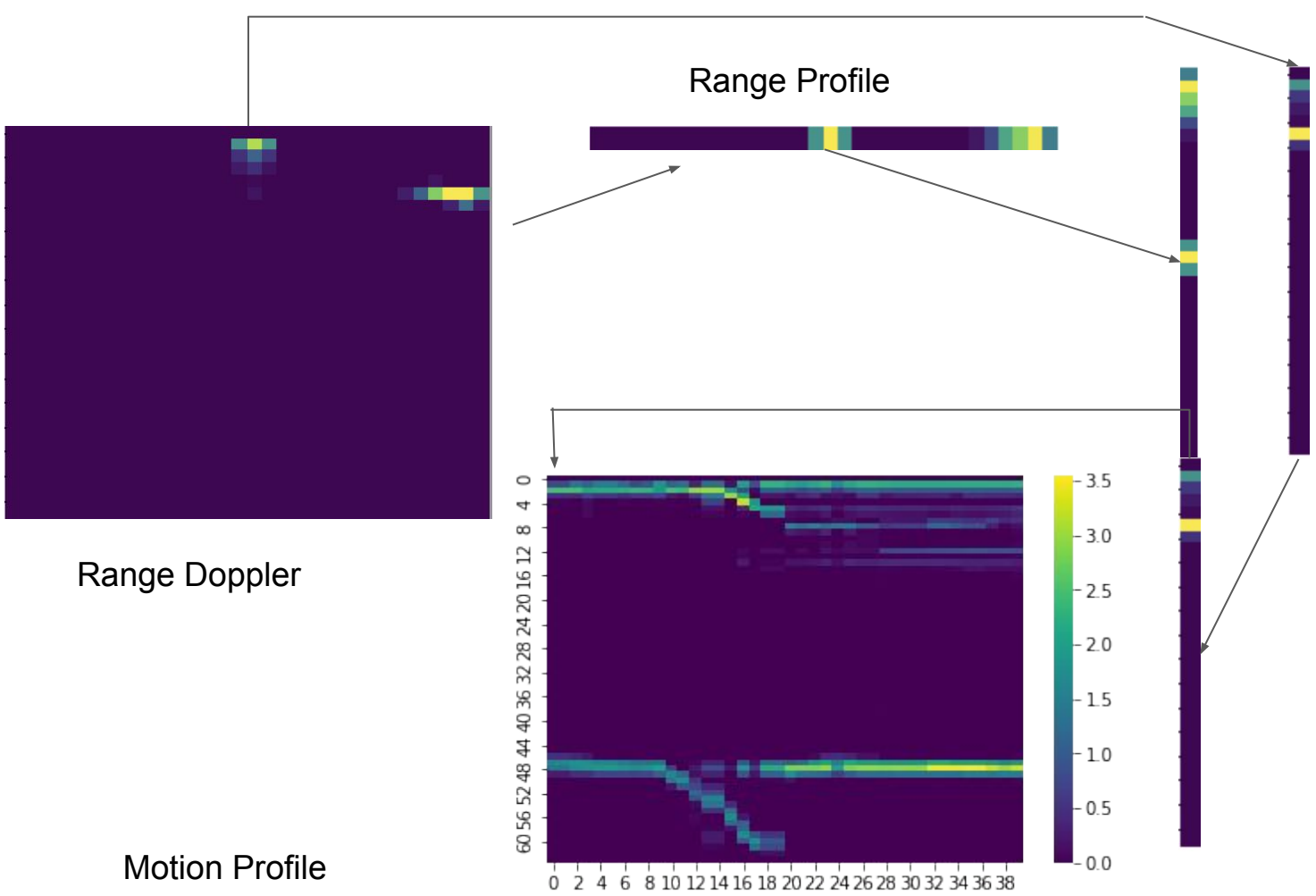
Background

$$RP_t^i = \sum_v RD^i(r, v, t)$$

$$DP_t^i = \sum_r RD^i(r, v, t)$$

Here,

RD is the range Doppler
RP is the range profile
DP is the doppler profile
i is receiver number or channel
t is the frame number
r is the range bin
v is the velocity bin
Motion Profile is a concatenation of Range Profile and Doppler Profile


Doppler Profile
Range Profile
Range Doppler
Motion Profile
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Experiments

Experimented with CNN architecture using variety of features. Used sub sampling and padding techniques to handle variation in length of inputs received.

Model Architecture: Modified LeNet[1]

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 6, 32, 32]	6,006
MaxPool2d-2	[-1, 6, 16, 16]	0
Conv2d-3	[-1, 12, 12, 12]	1,812
MaxPool2d-4	[-1, 12, 6, 6]	0
Linear-5	[-1, 216]	93,528
Linear-6	[-1, 54]	11,718
Linear-7	[-1, 11]	605

Total params: 113,669
Trainable params: 113,669
Non-trainable params: 0

[1]. X. zhang, Q. Wu and D. Zhao, "Dynamic Hand Gesture Recognition Using FMCW Radar Sensor for Driving Assistance," 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), 2018, pp. 1-6, doi: 10.1109/WCSP.2018.8555642.

Accuracy Metrics

Conducted robustness tests such as inter user testing, inter session testing
Pruned the data set based on the features which were not confusing and repeated the same tests(we removed gestures 3 and 10).
The results are as follows:

Feature	**Baseline**	**Leave One User Out (LOUO)**	**Leave One Session Out (LOSO)**	**Baseline [Pruned Data]**	**LOUO [Pruned Data]**	**LOSO [Pruned Data]**
Motion Profile	95.67%	90.18%	92.84%	98.12%	93.87%	97.95%
Sequence of RDIs	93.11%	86.22%	90.33%	96.86%	91.95%	96.58%

CNN Model Complexity

These statistics can help us know about the model complexity:
FLOP denotes Floating Point Operations during the inference stage.

Model	Time taken (ms)	FLOP (M)	Number of parameters
Motion Profile	0.89	3.91	766547
RDI	1.90	13.02	113669

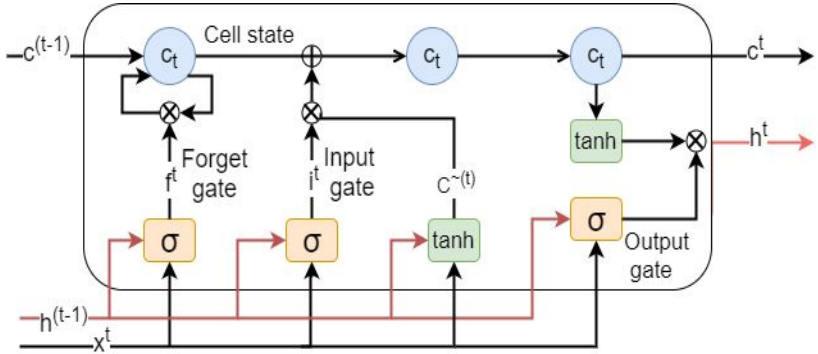
- These are calculated for a 40-frame gesture during the inference stage.
- The time taken is calculated on an i5, 1.8GHz machine.

BTP -1 Conclusions

- Literature review Summary:
  - Majority of the literature leave out details such as memory, power consumption, model robustness, prediction time
  - No standardised dataset large/diverse enough to be used as a common benchmark for testing these models
  - The variable sequence length of input data has not been emphasized
- CNN models are quick but are unable to handle variable size input and confuse between similar gestures (dataset pruning can cure this)
- We started to look at sequential models like LSTMs,Transformers etc

LSTM Encoder

- LSTM encoder is a model which can take a variable length sequence data and can return a fixed shape representation of our input.
- This representation of the input is encapsulated in “hidden state” of the model.
- Hidden state is basically a vector which is comprised of values from each of the LSTM cells, we get a hidden state from the model at every timestep.
- Typically the hidden state at the final time step of LSTM is used as the representation of input data



Source: researchgate.net/figure/Long-Short-Term-Memory-LSTM-cell-Fundamental-components-of-an-LSTM-cell-are-a-forget_fig1_329619613

Attention Mechanism

- Usage of the last hidden state as representation is not optimal as the model tends to forget some information over time
- Attention is a technique to get the representation of input by computing a weighted sum of states at all times where those weights are also learnable

Formally: $q^j = \max_{t \in (1 \ldots n)} (h_t^j)$ $\hat{h}_t = \frac{h_t}{||h_t||}$

$$\alpha_t = \frac{exp(\hat{h}_t^T q)}{\sum_{j=1}^{n} exp(\hat{h}_j^T q)}$$

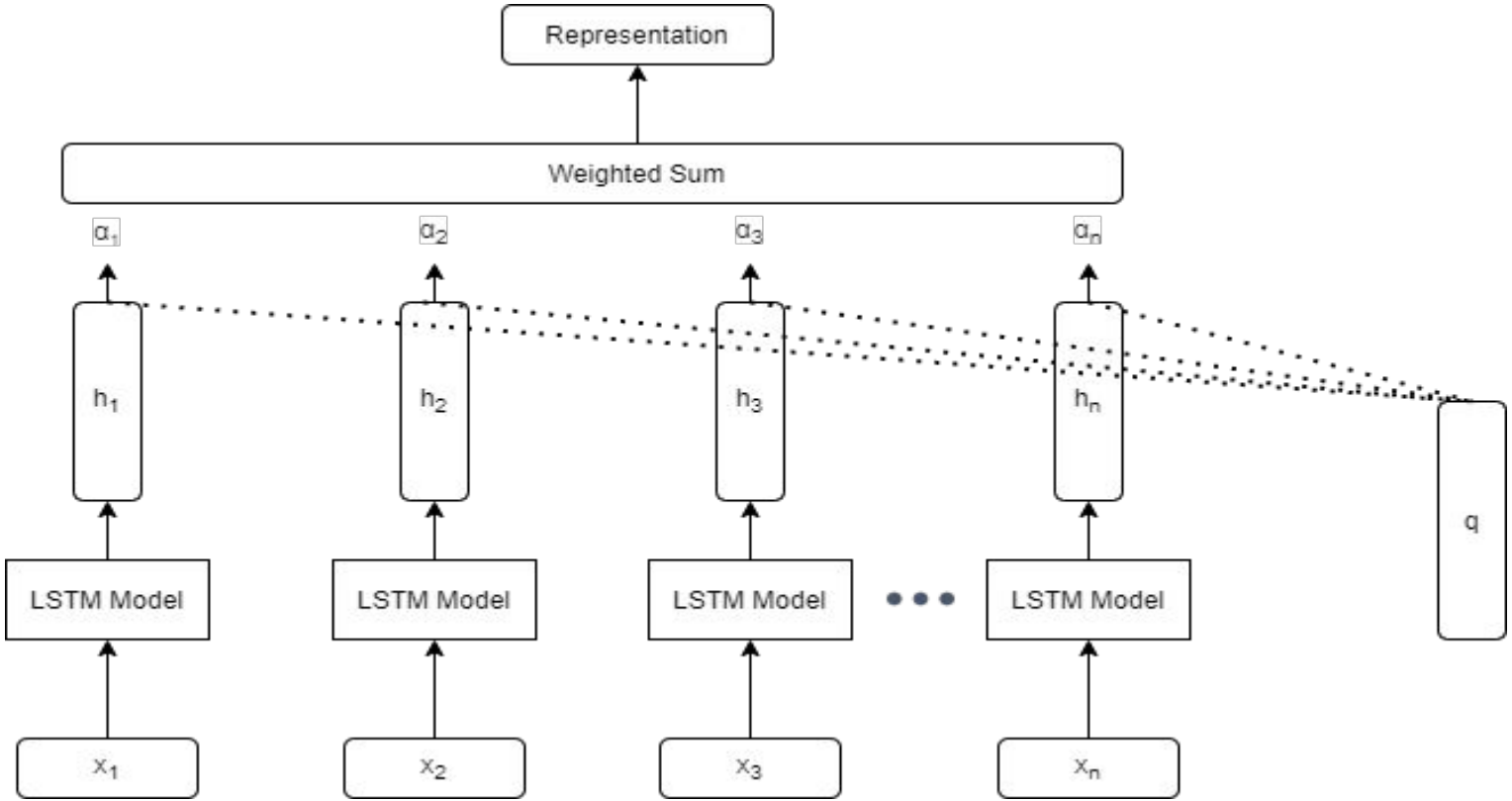
$$Representation = \sum_{t=1}^{n} \alpha_t h_t$$

Here,

j denotes the j^{th} dimension of h_t
t denotes the t^{th} time step
h_t denotes the hidden state at t^{th} time step

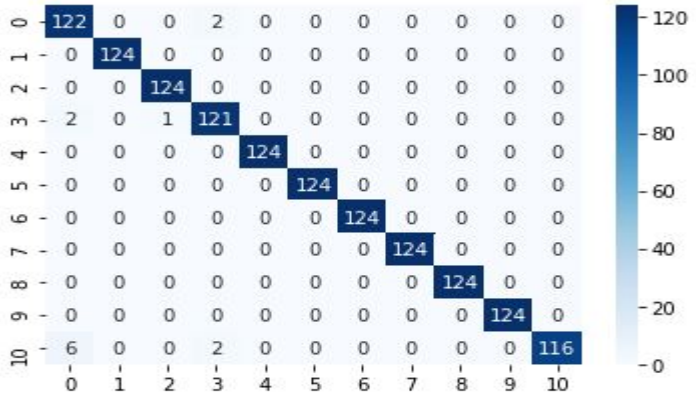
Source: P. Maini, K. Kolluru, D. Pruthi and Mausam . “Why and when should you pool? Analyzing Pooling in Recurrent Architectures” EMNLP (2020).
”

LSTM Architecture



LSTM Results

As we can see for this model the two bad performing gestures are also getting classified correctly around 94% of times which decreases the importance of pruning.



Confusion Matrix (2-layer model)

Model	**Baseline**	**Leave One User Out (LOUO)**	**Leave One Session Out (LOSO)**
2-layered	**99.04%**	92.07%	**96.25%**
1-layered	98.97%	**92.90%**	95.92%

LSTM Model Complexity

These statistics can help us know about the model complexity:
FLOP denotes Floating Point Operations during the inference stage.

Model	Time taken (ms)	FLOP (M)	Number of parameters
1-layered LSTM	4.91	15.82	202443
2-layered LSTM	9.29	31.58	334539

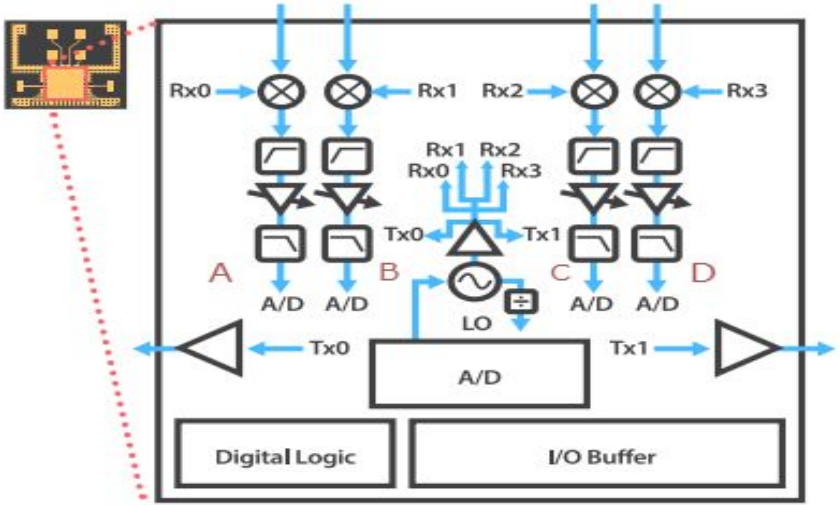
- These are calculated for a 40-frame gesture during the inference stage.
- The time taken is calculated on an i5, 1.8GHz machine.

Model	Baseline	LOUO	LOSO	Time taken(ms)	Number of parameters	FLOP (M)
GVLAD[2]	98.24%	91.38%	**97.75%**	4.8*	>10M	—
Res3DTENet[3]	96.99%	**92.25%**	—	—	—	—
LSTM 2-layered	**99.04%**	**92.07%**	96.25%	9.29	334539	31.58
LSTM 1-layered	**98.97%**	**92.90%**	95.92%	4.91	202443	15.82
CNN Motion Profile	95.67%	90.18%	92.84%	0.89	766547	3.91
	98.12%	93.87%	97.95%			
CNN Sequence of RDIs	93.11%	86.22%	90.33%	1.90	113669	13.02
	96.86%	91.95%	96.58%			

* time taken is calculated on an i7, 2.2 GHz machine.

SOLI CHIP(Labelled Channel wise)

- Channel Refinement : Do we need all 4 channels or can we reduce channels without significant loss in accuracy (In consideration for edge devices)



Source: Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: ubiquitous gesture sensing with millimeter wave radar. ACM Trans. Graph. 35, 4, Article 142 (July 2016), 19 pages. DOI:https://doi.org/10.1145/2897824.2925953

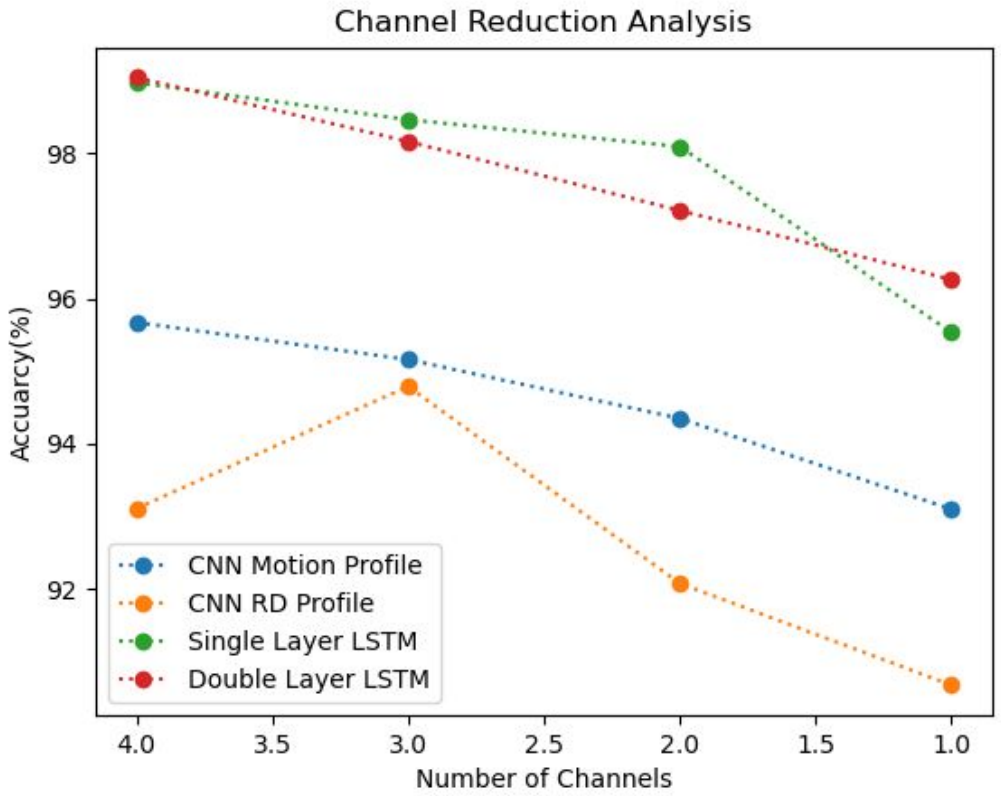
Channel Reduction Analysis

Model	Number of channels used	Accuracy (50-50 split)
CNN_MP	4(a,b,c,d)	95.67%
CNN_MP	3(a,b,c)	95.16%
CNN_MP	2(a,b)	94.35%
CNN_MP	1(a)	93.10%
CNN_RDI_INP	4(a,b,c,d)	93.11%
CNN_RDI_INP	3(a,b,c)	94.79%
CNN_RDI_INP	2(a,b)	92.08%
CNN_RDI_INP	1(a)	90.68%

Channel Reduction Analysis

Model	Number of channels used	Accuracy (50-50 split)
LSTM (1 layer)	4(a,b,c,d)	98.97%
LSTM (1 layer)	3(a,b,c)	98.46%
LSTM (1 layer)	2(a,b)	98.09%
LSTM (1 layer)	1(a)	95.53%
LSTM (2 layer)	4(a,b,c,d)	99.04%
LSTM (2 layer)	3(a,b,c)	98.16%
LSTM (2 layer)	2(a,b)	97.21%
LSTM (2 layer)	1(a)	96.26%

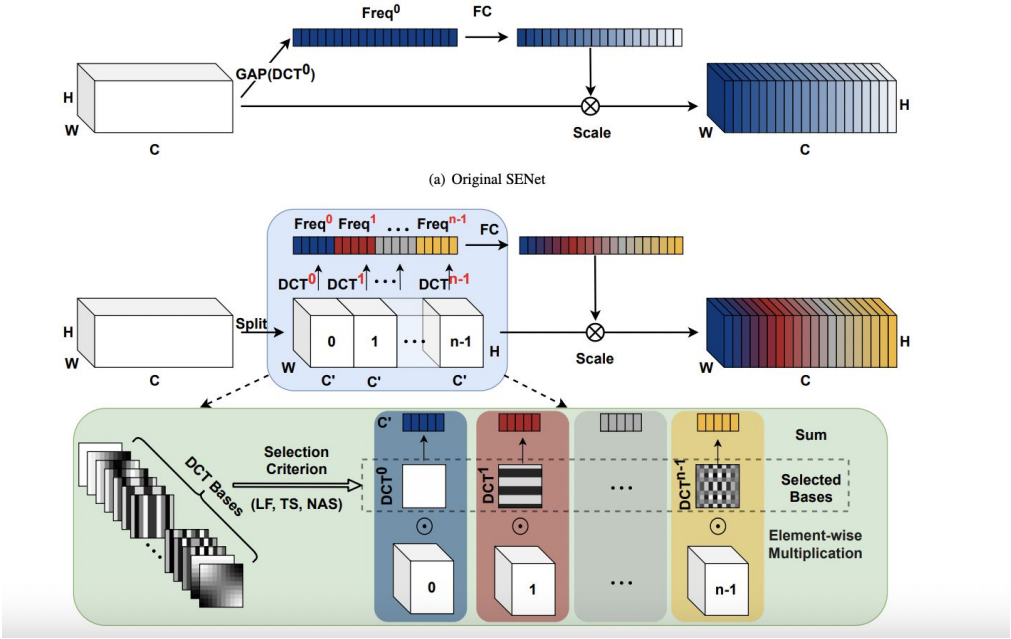
Impact of Number of channels



Learnings

- Denser Hand crafted features are much better than Vanilla Range Dopplers
- CNN can work reasonably well for edge devices with lower memory requirements and pruned gesture set
- Sequential Models can not only handle variable length data but also perform better leading to increased accuracy in all metrics
- Attention Mechanism boosted LSTM accuracy. Can we use state of the art attention mechanisms used in CNN and Transformers and build those models for accuracy boost ?
- Can these complex models be used on Edge Devices?

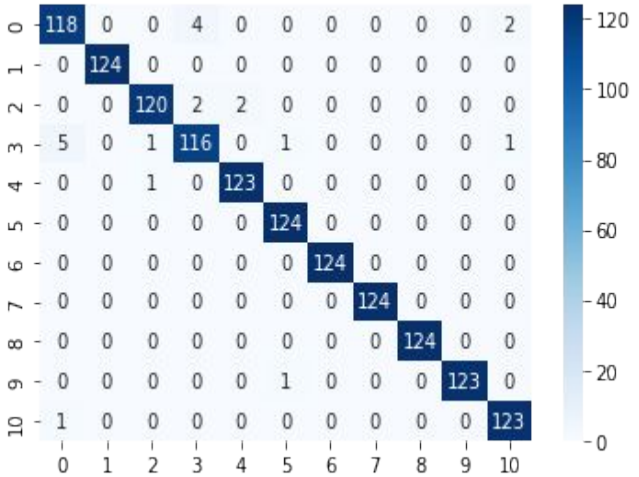
Attention in CNN: FcaNet



Source: Qin, Z., Zhang, P., Wu, F., & Li, X.. (2020). FcaNet: Frequency Channel Attention Networks.

Performance on Soli

As we can see for this model the two bad performing gestures are also getting classified correctly which decreases the importance of pruning making State Of The Art (SOTA) Attention based CNN architectures a viable option



Model	**Baseline**
FCA based Alex Net	**98.46%**

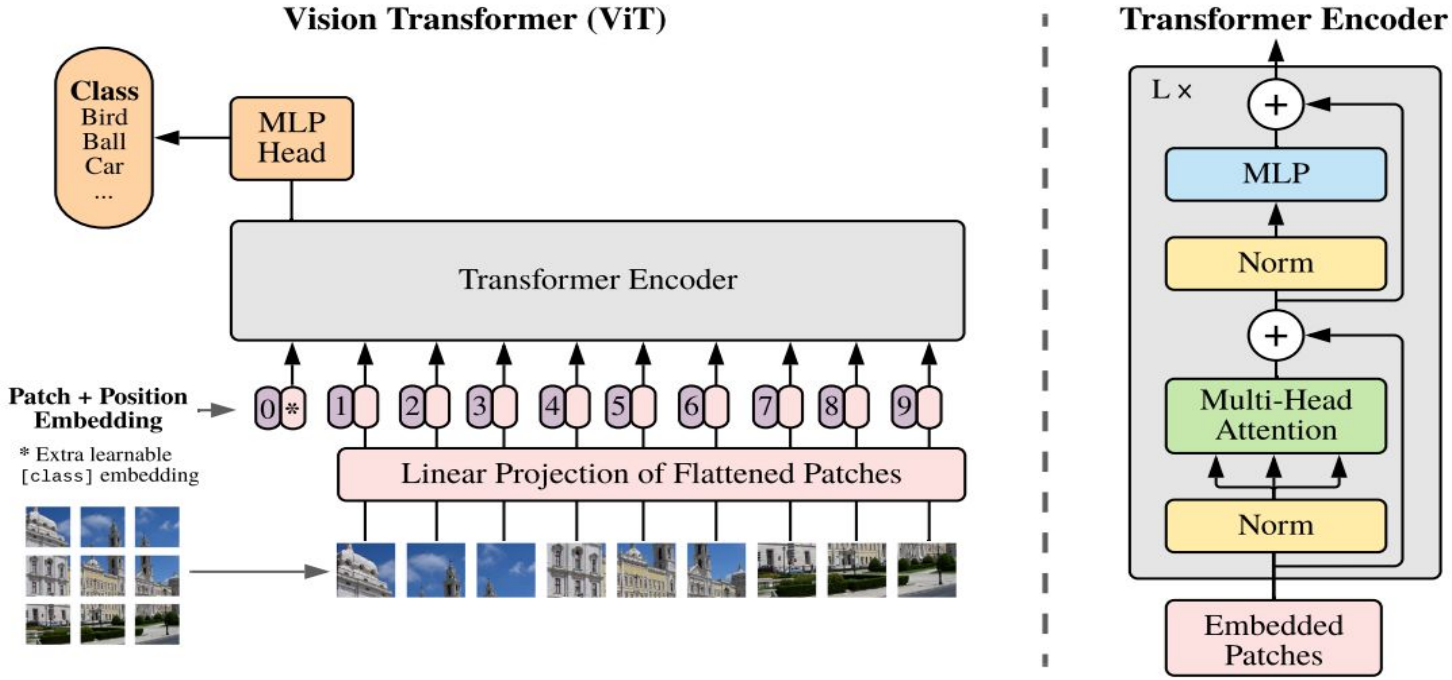
Model Complexity: FcaNet

These statistics can help us know about the model complexity:
FLOP denotes Floating Point Operations during the inference stage.

Model	Time taken (ms)	FLOP (M)	Number of parameters
Motion Profile	51.5	120	57.1 M

- These are calculated for a 40-frame gesture during the inference stage.
- The time taken is calculated on an i5, 1.8GHz machine.

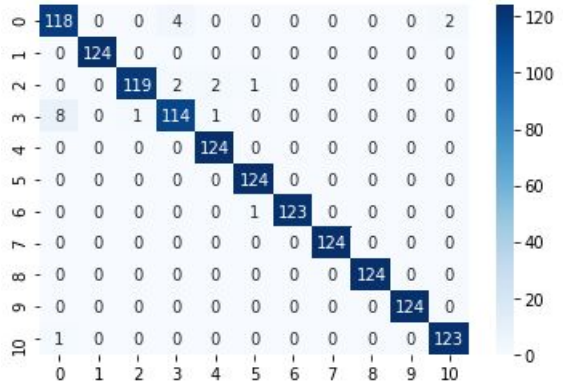
Vision Transformer



Source: Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*

Performance on Soli

For ViT as well the two hardest to classify gestures are Being identified pretty well. In the use case where Number of gestures are large and very high accuracies Are needed ViT can be considered.



Model	**Baseline**
ViT-Base (12 Blocks)	**98.31%**

Model Complexity: ViT

These statistics can help us know about the model complexity:
FLOP denotes Floating Point Operations during the inference stage.

Model	Time taken (ms)	FLOP	Number of parameters
Motion Profile	347.34	17 B	85.6 M

- These are calculated for a 40-frame gesture during the inference stage.
- The time taken is calculated on an i5, 1.8GHz machine.

Review of Models Used

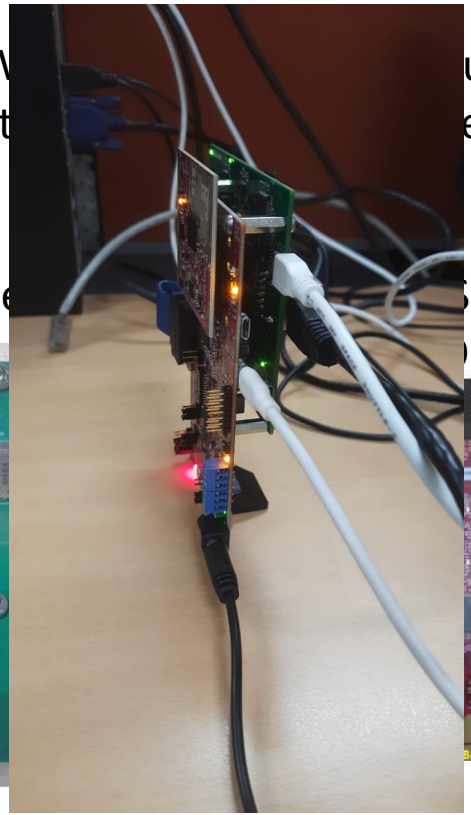
Model	Accuracy (%)	FLOP	Time Taken (ms)	Number of Parameters
CNN	95.67	3.91 M	0.89	766547
LSTM	99.04	31.58 M	9.29	334539
FcaNet	98.46	120 M	51.5	57.1 M
ViT	98.31	17 B	347.34	85.6 M

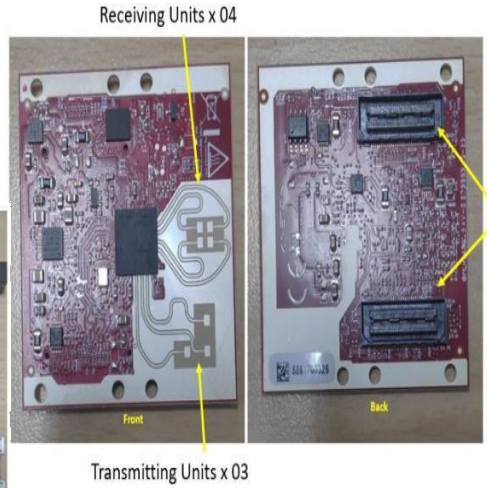
Radar Setup

We used a Texas Instruments mmW ur work. This hardware platform is based on three elect elow and their details are given in the following sections.

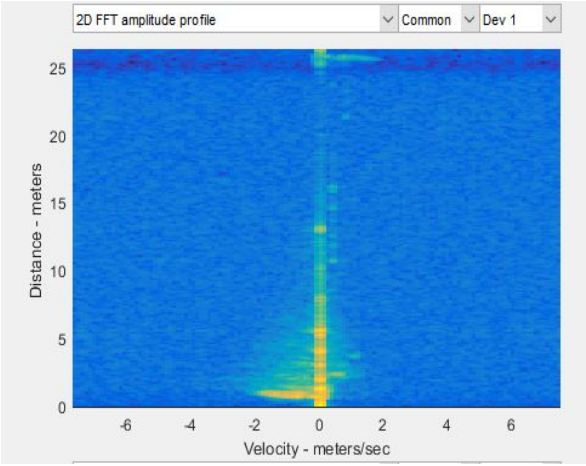
- mmWave overhead detection se)

OST

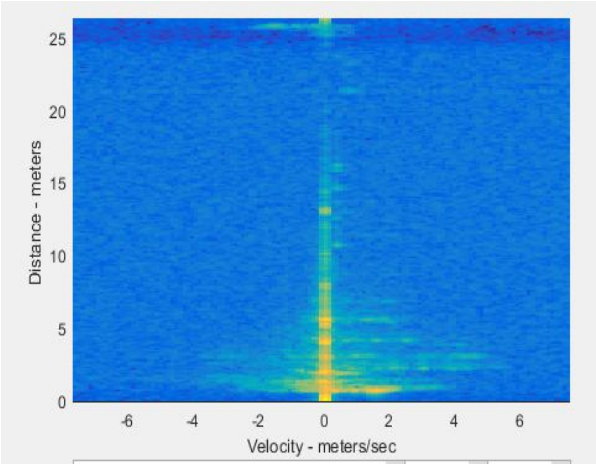




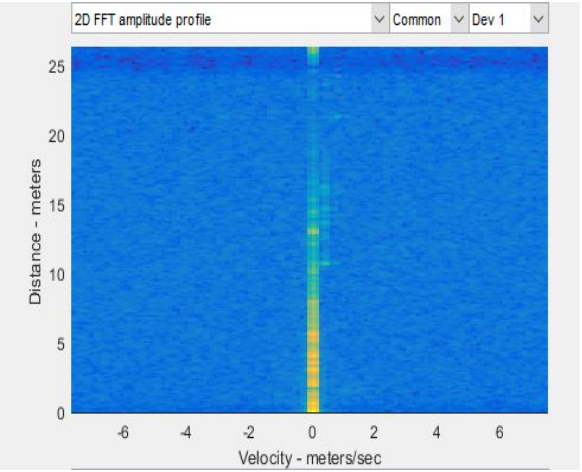
Coarse Motion Detection



Hand moving away from the radar

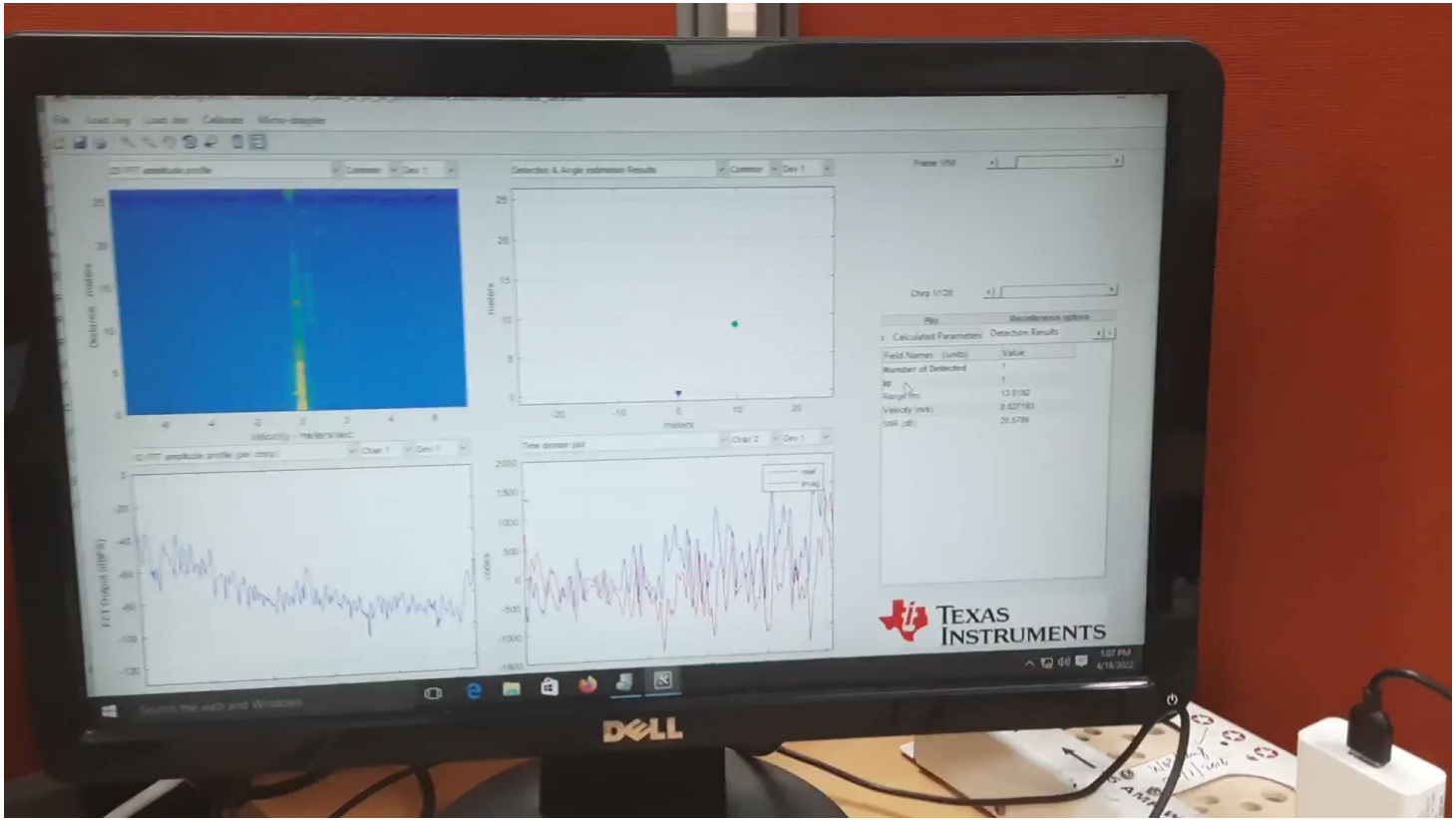


Hand moving towards the the radar



Hand still

Data Sample Demo



Project Outcomes

- Features such as Motion Profile are performing well instead of the complete RDIs which are very sparse
- Models such as CNN, LSTM combined with handcrafted features such as motion profile are good candidates for our target devices.
- SOTA attention based models such as ViT, FcaNet also give very good accuracy but have very high memory and compute requirements.
- Coarse motion has been successfully captured using real hardware

Future Targets

- Data set collection on TI provided radars and training cum testing of models on dataset for selection of the best performing model
- Implementing the gesture recognition pipeline on actual hardware for demo
- Conducting real-time, power and robustness tests [such as including left handed users, varying hand sizes]

References

1. X. zhang, Q. Wu and D. Zhao, "Dynamic Hand Gesture Recognition Using FMCW Radar Sensor for Driving Assistance," 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), 2018, pp. 1-6, doi: 10.1109/WCSP.2018.8555642.
2. A. D. Berenguer, M. C. Oveneke, H. -U. -R. Khalid, M. Alioscha-Perez, A. Bourdoux and H. Sahli, "GestureVLAD: Combining Unsupervised Features Representation and Spatio-Temporal Aggregation for Doppler-Radar Gesture Recognition," in IEEE Access, vol. 7, pp. 137122-137135, 2019, doi: 10.1109/ACCESS.2019.2942305.
3. Range-Doppler Hand Gesture Recognition Using Deep Residual-3DCNN with Transformer Network. G Jaswal, S Srirangarajan, SD Roy - International Conference on Pattern Recognition, 2021

THANK YOU

Questions ?