Radar based Gesture Recognition

Lakshya Kumar Tangri 2018EE10222
Somanshu Singla 2018EE10314

Faculty Supervisor: Prof. Seshan Srirangarajan

15 December 2021

Introduction

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

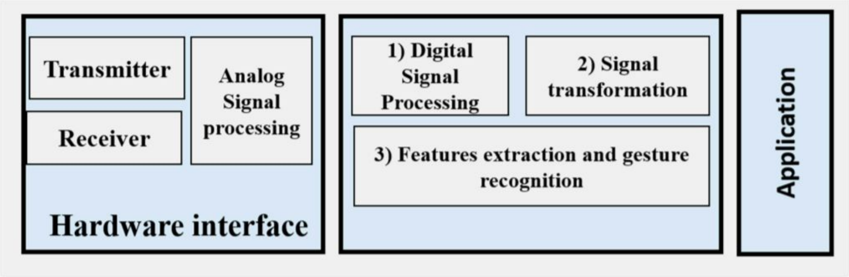
Short-range radars have the ability to detect micro-movements with high precision and accuracy making it a good candidate for realizing gesture-based interfaces

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

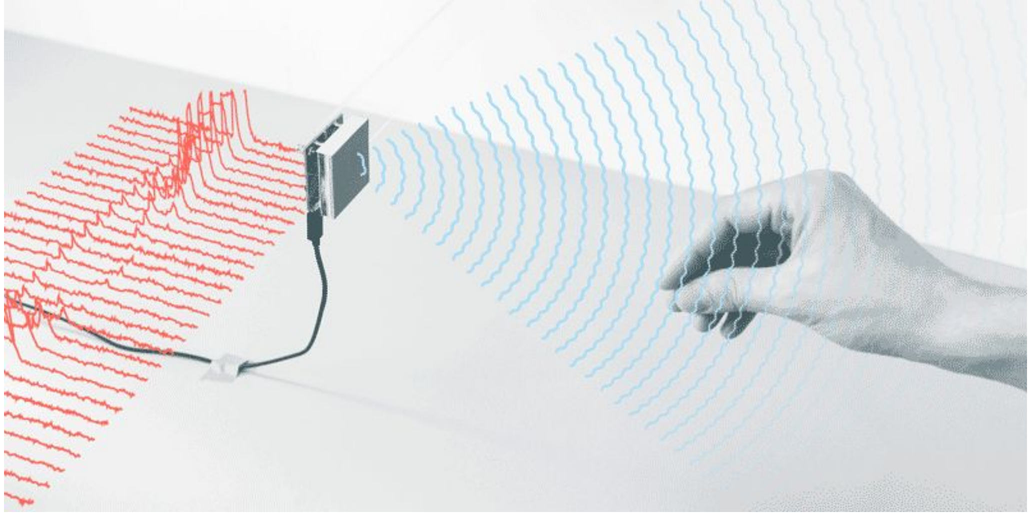
To create a robust, radar based power efficient system which can classify predefined gestures accurately in real time on a mobile device

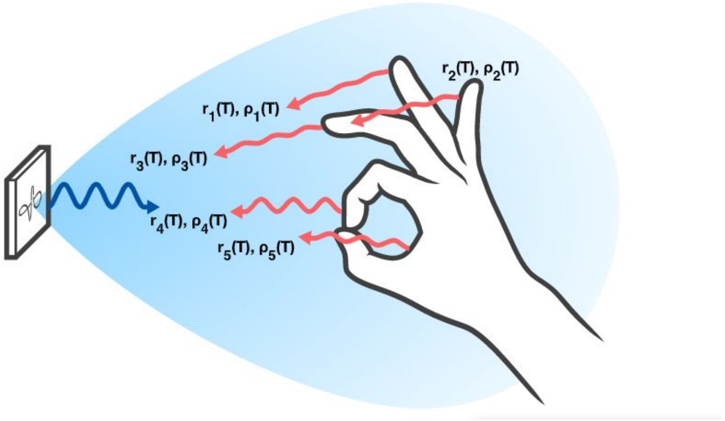
- Robustness: With respect to different users(inter-user) and different gesture instances of a single user(inter-session)
- Accuracy: To be able to correctly classify the input gesture
- Power Efficiency and Real Time: Ability to recognize gesture quickly for a seamless user experience without draining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

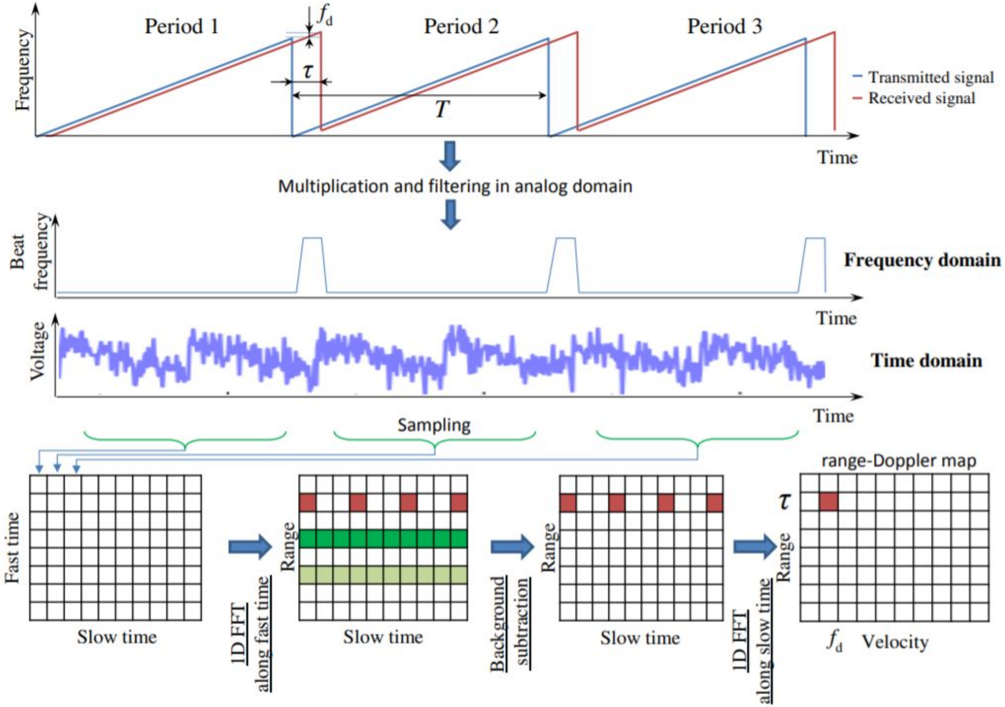
Working

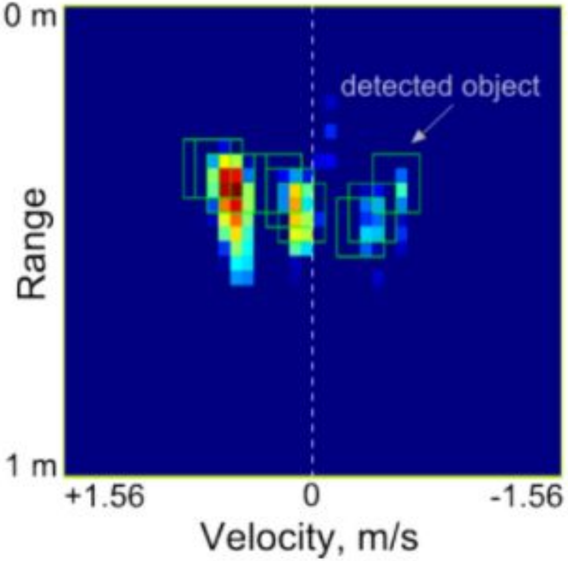




Source: Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: ubiquitous gesture sensing with millimeter wave radar. ACM Trans. Graph. 35, 4, Article 142 (July 2016), 19 pages. DOI:https://doi.org/10.1145/2897824.2925953

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.

Literature Survey

Types of papers:

I. Analyzing properties of FMCW Radar
II. Traditional machine learning
III. Deep Learning

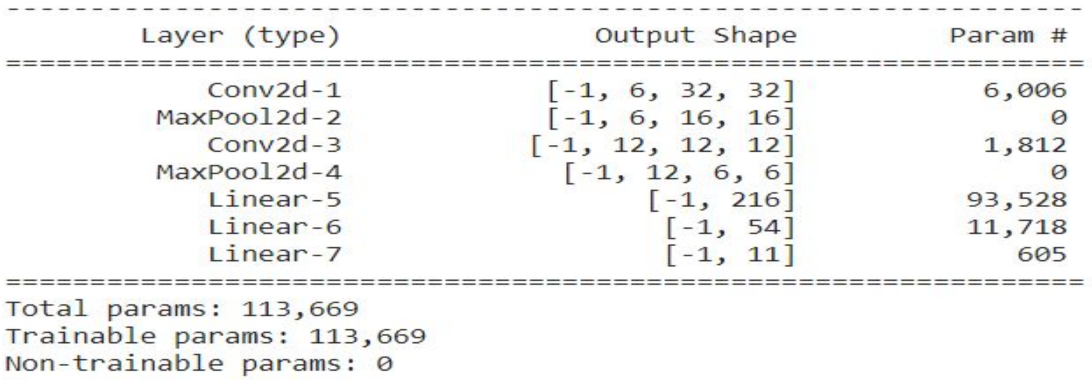
Observations:

- 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

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[7]



```
----------------------------------------------------------------
        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
```

[7]. 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.

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

Experiment Results

Compared accuracy metrics of some basic features:
All the accuracy metrics reported are for 50-50 Train Test Split.

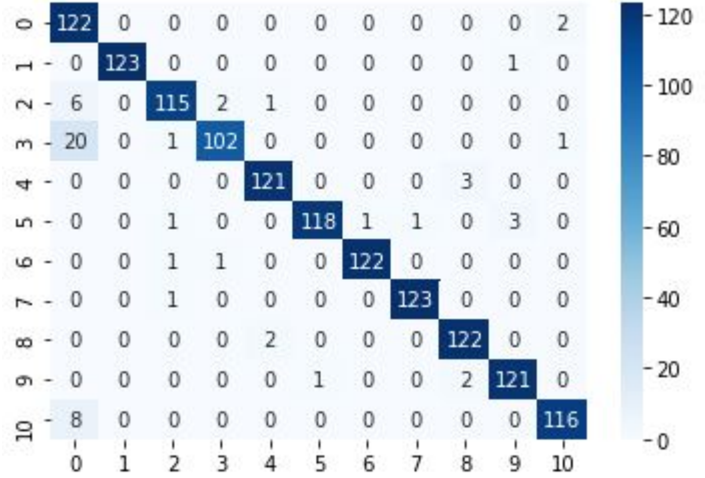
By using Range profile as feature:
Accuracy = *84.897%*

By using Doppler profile as feature:
Accuracy = *89.589%*

By using Motion Profile feature:
Accuracy = *95.45%*

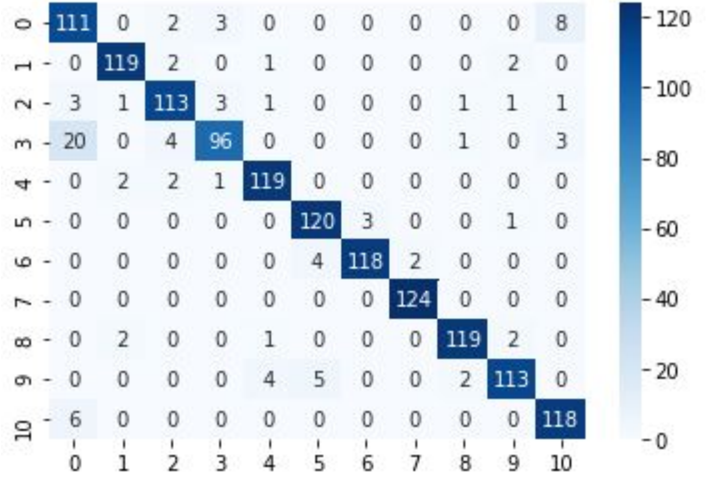
By using complete sequence of RDIs:
Accuracy = *93.40%*

Confusion Matrix on Unpruned Data Set



	0	1	2	3	4	5	6	7	8	9	10
0	122	0	0	0	0	0	0	0	0	0	2
1	0	123	0	0	0	0	0	0	0	1	0
2	6	0	115	2	1	0	0	0	0	0	0
3	20	0	1	102	0	0	0	0	0	0	1
4	0	0	0	0	121	0	0	0	3	0	0
5	0	0	1	0	0	118	1	1	0	3	0
6	0	0	1	1	0	0	122	0	0	0	0
7	0	0	1	0	0	0	0	123	0	0	0
8	0	0	0	0	2	0	0	0	122	0	0
9	0	0	0	0	0	1	0	0	2	121	0
10	8	0	0	0	0	0	0	0	0	0	116

For Motion Profile feature



	0	1	2	3	4	5	6	7	8	9	10
0	111	0	2	3	0	0	0	0	0	0	8
1	0	119	2	0	1	0	0	0	0	2	0
2	3	1	113	3	1	0	0	0	1	1	1
3	20	0	4	96	0	0	0	0	1	0	3
4	0	2	2	1	119	0	0	0	0	0	0
5	0	0	0	0	0	120	3	0	0	1	0
6	0	0	0	0	0	4	118	2	0	0	0
7	0	0	0	0	0	0	0	124	0	0	0
8	0	2	0	0	1	0	0	0	119	2	0
9	0	0	0	0	4	5	0	0	2	113	0
10	6	0	0	0	0	0	0	0	0	0	118

For Sequence of RDIs

Gesture Set Pruning

- Observed that features mentioned in the literature don’t do well on the complete gesture set and after viewing the confusion matrices concluded that out of all gestures, model’s performance is not upto the mark for some gestures(i.e it confuses between two similar gestures) and good for others.

- In a practical setting we don’t really need a gesture set containing 10- 15 gestures even some 5 - 7 gestures (being ergonomic) can suffice , so keeping this in mind we test our model again on removing of these bad performing gestures

[1] 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

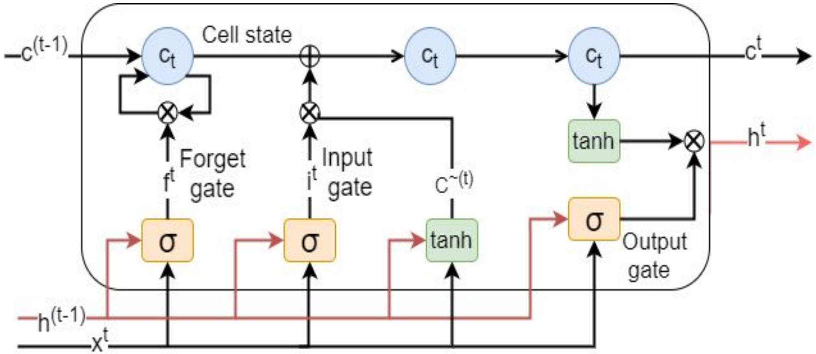
Accuracy Metrics

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	95.45%	90.18%	92.836%	98.656%	93.8%	97.95%
Sequence of RDIs	93.40%	86.218%	90.327%	96.864%	91.95%	96.578%

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:

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

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

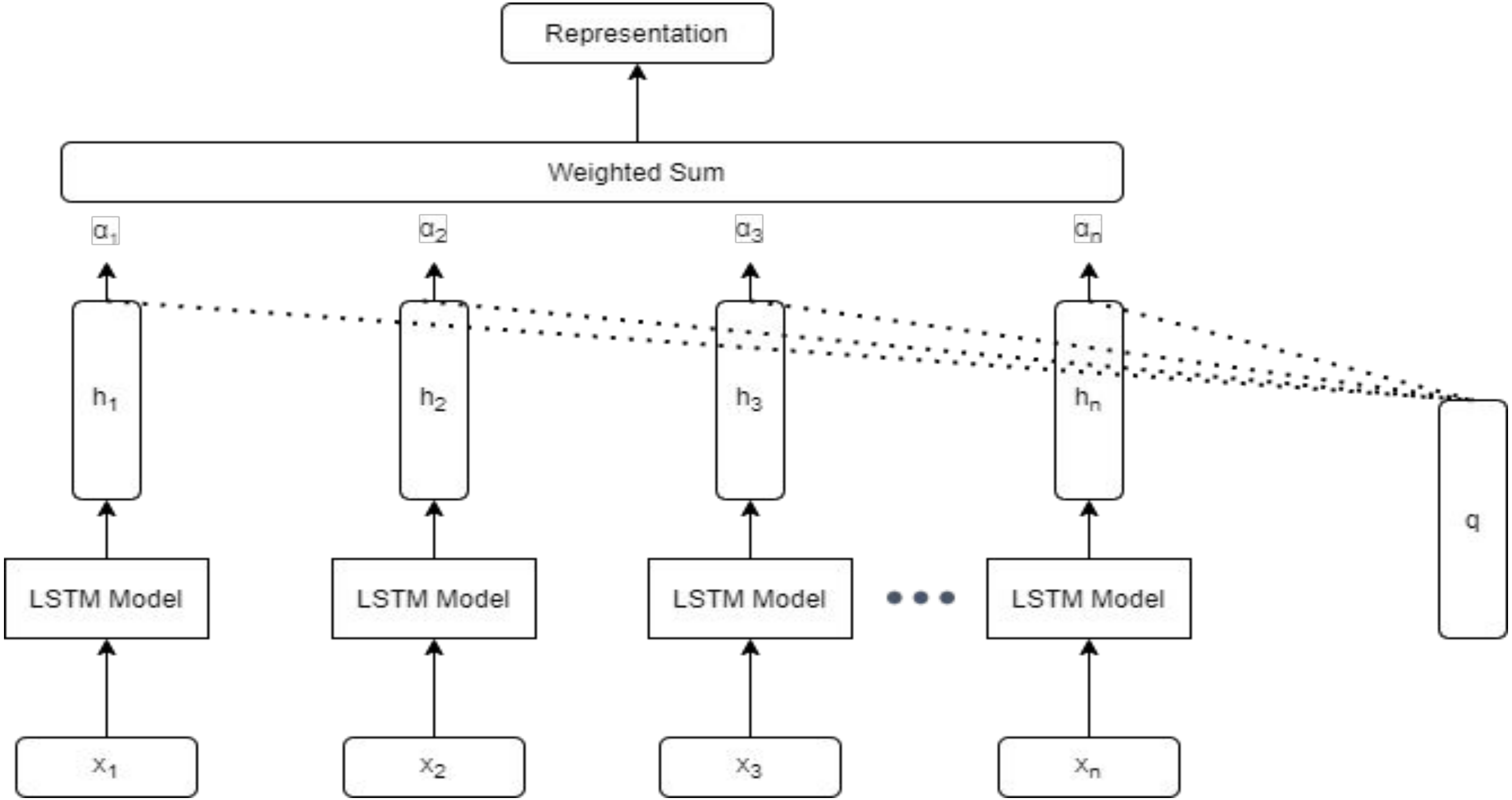
$$Representation = \sum_{t=1}^{n} \alpha_t h_t$$

Here,

j denotes the jth dimension of h$_t$
t denotes the tth time step
h$_t$ denotes the hidden state at tth time step

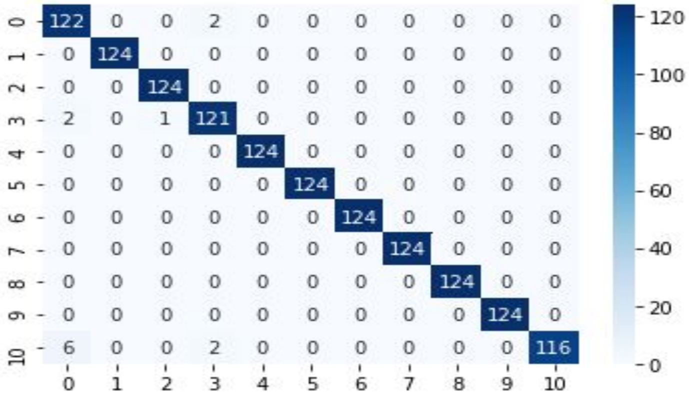
Source: P. Maini, K. Kolluru, D. Pruthi and Mausam . “Why and when should you pool? Analyzing Pooling in Recurrent Architectures” EMNLP (2020).
”

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.



	0	1	2	3	4	5	6	7	8	9	10
0	122	0	0	2	0	0	0	0	0	0	0
1	0	124	0	0	0	0	0	0	0	0	0
2	0	0	124	0	0	0	0	0	0	0	0
3	2	0	1	121	0	0	0	0	0	0	0
4	0	0	0	0	124	0	0	0	0	0	0
5	0	0	0	0	0	124	0	0	0	0	0
6	0	0	0	0	0	0	124	0	0	0	0
7	0	0	0	0	0	0	0	124	0	0	0
8	0	0	0	0	0	0	0	0	124	0	0
9	0	0	0	0	0	0	0	0	0	124	0
10	6	0	0	2	0	0	0	0	0	0	116

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%

Progress Made

- Experimented with various features and architectures proposed in the literature using the publicly available Google Soli Data Set[6]
- Realised that even though the input data can have variable length samples, the current work in this area doesn't really delve deep and instead uses stopgap approaches.
- Experimented with attention based sequence architecture to handle variable length samples

[6] Wang, S.; Song, J.; Lien, J.; Poupyrev, I.; Hilliges, O. Interacting with soli: Exploring fine-grained dynamic gesture recognition inthe radio-frequency spectrum. In Proceedings of the 29th Annual Symposium on User Interface

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
- Conducting real-time, power and robustness tests [such as including left handed users, varying hand sizes]

THANK YOU

Questions ?