Mobile Sensing: Deep Learning on Mobiles

Master studies, Spring 2021/2022

Dr. Octavian Machidon
octavian.machidon@fri.uni-lj.si

University of Ljubljana
Faculty of Computer and Information Science

Based on lecture slides by dr. Veljko Pejović

Lecture outline

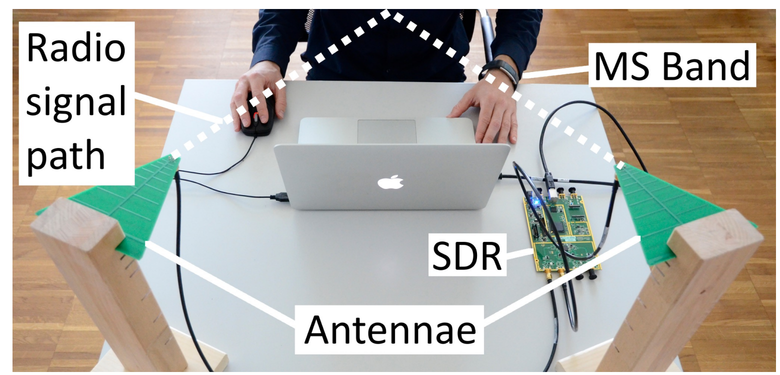
- Deep Learning in Mobile Sensing
  - Applications
  - Key benefits
  - Challenges
  - Solutions for reducing the cost of Deep Learning

Deep Learning in Mobile Sensing

- Mobile computer vision
  - Object recognition with CNN
- Activity recognition
  - CNNs and RNNs
  - *"Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges"* by Nweke et al.

Deep Learning in Mobile Sensing

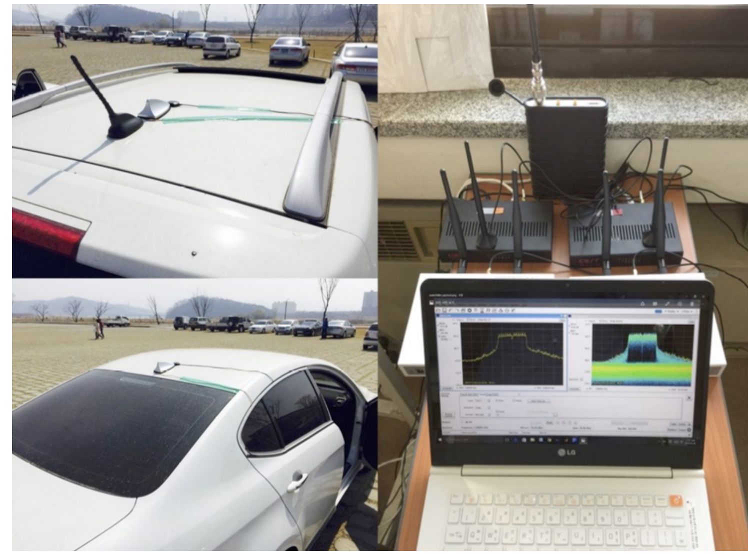
- Depression detection
  - Autoencoder for relevant mobility feature identification
  - "*Using autoencoders to automatically extract mobility features for predicting depressive states*" by Mehrotra et al.
- Cognitive load inference
  - LSTM processing wireless radar signals reflected off a person's chest (i.e. breathing, heartbeats)



T. Matkovic and V. Pejovic, *Wi-Mind: Wireless Mental Effort Inference*
Ubittention workshop with UbiComp'18, Singapore, October 2018.

Deep Learning in Mobile Sensing

- Predicting wireless signal quality in vehicular communication
  - LSTM on wireless spectrum sensing data

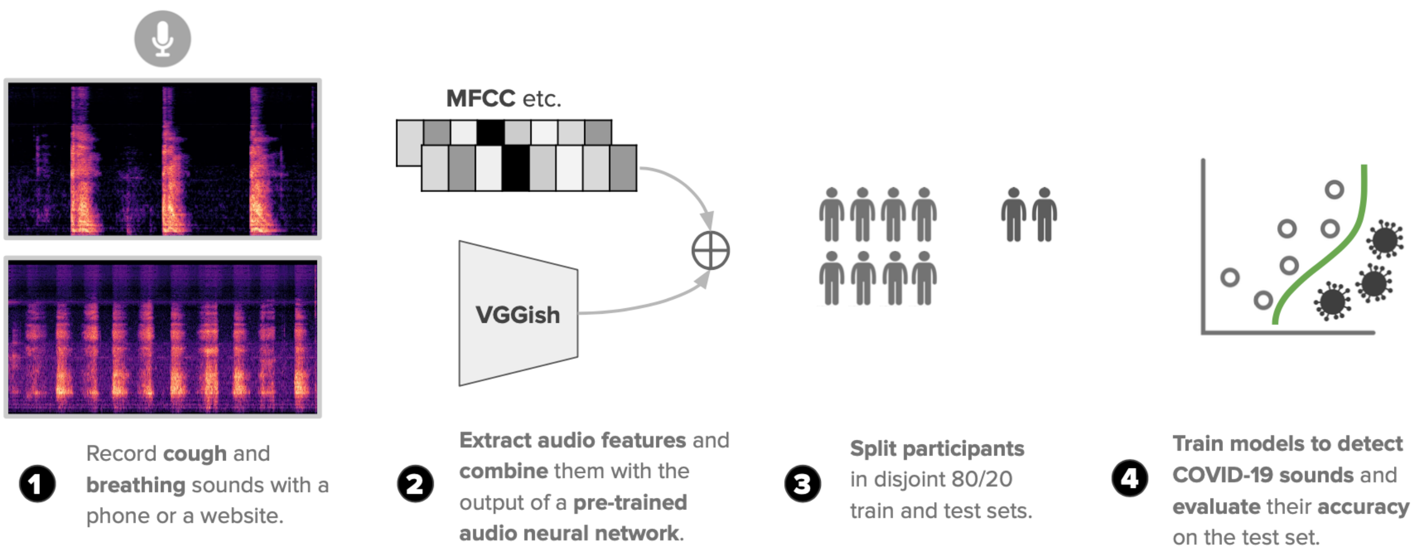


University of Ljubljana
Faculty of Computer and Information Science

J. Joo, M.C. Park, D.S. Han, and V. Pejovic
Deep Learning-based Channel Prediction in Realistic Vehicular Communications, IEEE Access (2019).

Deep Learning in Mobile Sensing

- Healthcare
  - Predict an onset of a disease



University of Ljubljana
Faculty of Computer and Information Science

Brown, Chloë, et al.
Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data
arXiv:2006.05919 (2020).

Deep Learning in Mobile Sensing

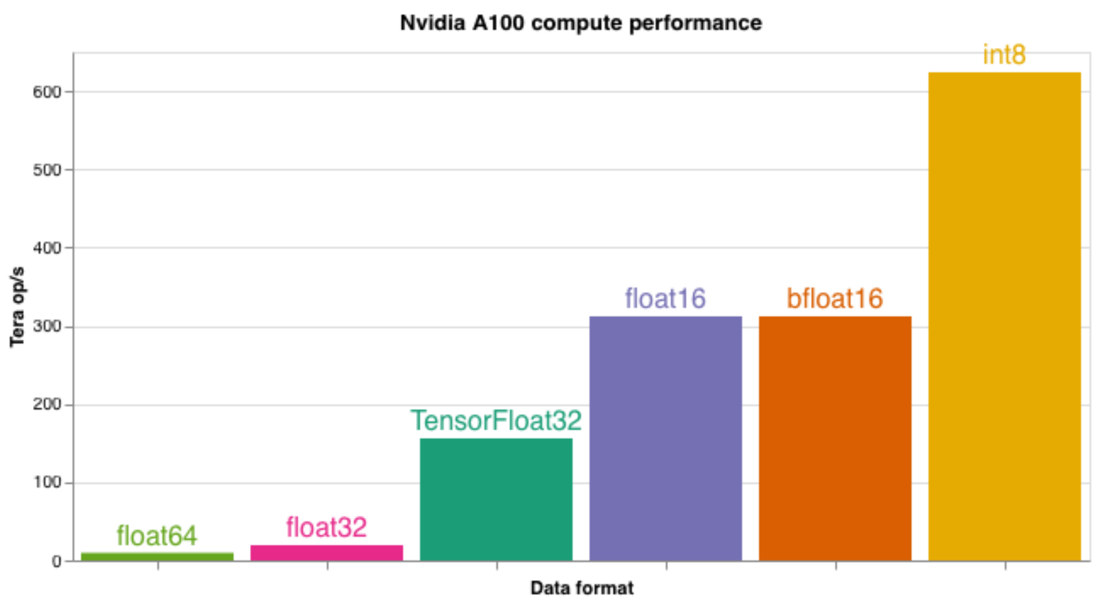
- Key benefits of deep learning on mobile devices
  - Reduced delay – inferences can happen faster, if we don't need to send the data back and forth to the server
  - Reduction in bandwidth usage – spectrum is a limited resource
  - Operation when the connectivity is not available
  - Privacy – keeping the data (your photos, voice recordings) on the device

Challenges with Deep Learning on Mobiles

- A neural network requires:
  - A lot of storage space/mem – millions of parameters
  - A lot of training data – NNs tend to work well only when there is a lot of (labelled) data available
  - A lot of computation – matrix multiplication, backpropagation, batch training, etc.
- Solving the problems:
  - Training data – background sensing, semi-supervised learning, crowdsourced efforts (Google)
  - Computation and storage space/mem – we cannot rely on the next generation of devices to solve everything

Reducing the Cost of Deep Learning

- Quantization:
  – Instead of 32b floats, use 16b floats, 8b/4b/2b integers, or binary (1/0) weights and activations
  – It saves not just storage, but also computational time



Source: https://mlech26l.github.io/pages/jupyter/quantization/2020/06/28/quantization.html

University of Ljubljana
Faculty of Computer and Information Science

Example from towardsdatascience.com/how-to-accelerate-and-compress-neural-networks-with-quantization-edfbbabb6af7

Reducing the Cost of Deep Learning

- Quantization example (from float32 to int8):
  - suppose weights and activations are in the range [-a, a)
  - scale the output to [-128, 128): $x \mapsto \left\lfloor 128 \frac{x}{a} \right\rfloor.$
  - e.g.

Range [-1,1)

$$\begin{pmatrix} -0.18120981 & -0.29043840 \\ 0.49722983 & 0.22141714 \end{pmatrix} \begin{pmatrix} 0.77412377 \\ 0.49299395 \end{pmatrix} = \begin{pmatrix} -0.28346319 \\ 0.49407474 \end{pmatrix}$$

- quantize:

not really int8, though

$$\begin{pmatrix} -24 & -38 \\ 63 & 28 \end{pmatrix} \begin{pmatrix} 99 \\ 63 \end{pmatrix} = \begin{pmatrix} -4770 \\ 8001 \end{pmatrix}.$$

- dequantize: $x \mapsto \frac{ax}{16384}$ we get: $\begin{pmatrix} -0.2911377 \\ 0.48834229 \end{pmatrix}.$

University of Ljubljana
Faculty of Computer and Information Science

Reducing the Cost of Deep Learning

- Benefits of model quantization for selected CNN models:

Model	Acc (orig.)	Acc (quant.)	Latency (orig) (ms)	Latency (quant.) (ms)	Size (orig.) (MB)	Size (quant.) (MB)
Mobilenet-v1	0.709	0.657	124	112	16.9	4.3
Mobilenet-v2	0.719	0.637	89	98	14	3.6
Inception_v3	0.780	0.772	1130	845	95.7	23.9
Resnet_v2	0.770	0.768	3973	2868	178.3	44.9

Source: https://www.tensorflow.org/lite/performance/model_optimization

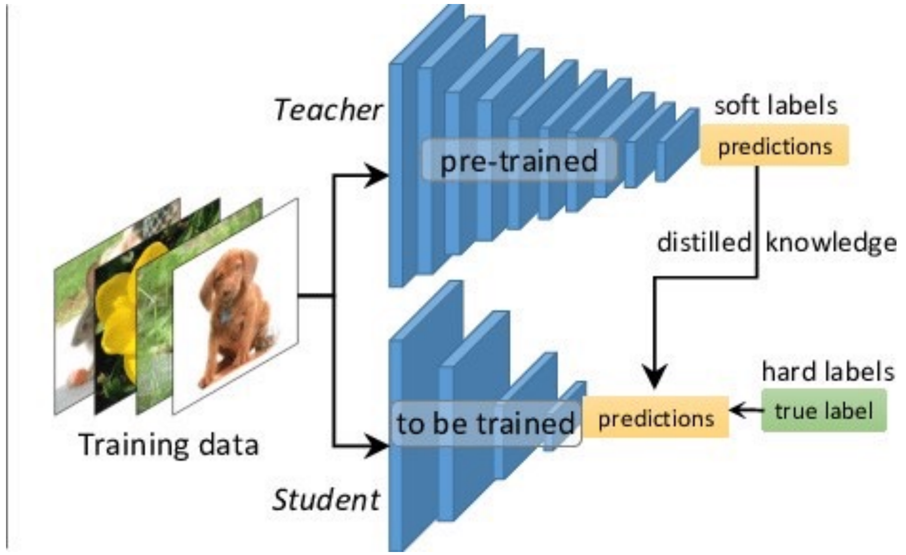
Reducing the Cost of Deep Learning

- Weight sharing/virtualization:
  - Represent multiple similar weights with a single value (often in combination with quantization)
- Pruning:
  - Reduce the number of weights after the training
  - Prune weights or whole neurons (unstructured pruning) vs. prune CNN filters or channels (structured pruning)
  - How to select what to prune?
- Matrix decomposition:
  - Use Singular Value Decomposition (SVD) to replace an $m \times n$ matrix with two smaller matrices of sizes $m \times c$ and $c \times n$
    - Total calculation drops from $O(m \times n)$ to $O(c \times (m+n))$

$c \ll m, n$

Reducing the Cost of Deep Learning

- Other, more advanced approaches
  - Knowledge distillation
    - A larger Teacher network "transfers" knowledge to a smaller Student network



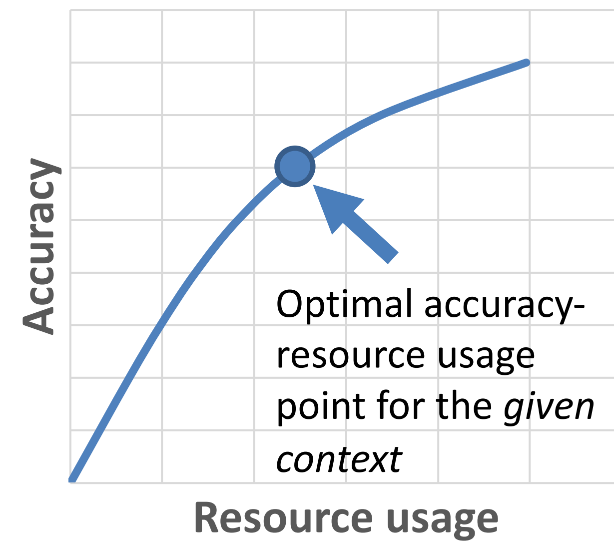
Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv:1503.02531*.

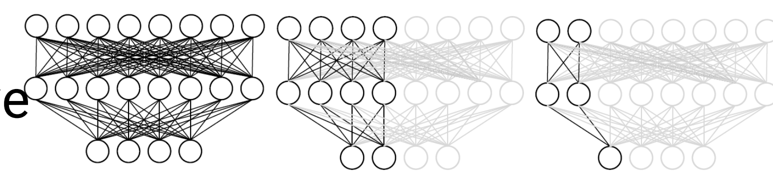
Reducing the Cost of Deep Learning

- Common drawbacks of these techniques:
  - Permanently ”impaired” network
  - Single point on the accuracy vs. resource usage trad—off curve
  - In mobile sensing the context varies, and so do the accuracy requirements and the available resources
- Solution: dynamic network compression techniques

Reducing the Cost of Deep Learning

- How to dynamically adjust to the context?
  - Dynamic quantization
  - Dynamic pruning
- More advanced approaches:
  - Slimmable Neural Networks
    - The same model can run at different widths allowing adaptive accuracy-efficiency trade-off
    - SNN for HAR





Machidon, O., et al. Queen Jane Approximately: Enabling Efficient Neural Network Inference with Context-Adaptivity, *EuroMLSys '21*

University of Ljubljana
Faculty of Computer and Information Science

Yu, Jiahui, et al. "Slimmable neural networks." *arXiv:1812.08928* (2018).

Mobile Sensing:
Deep Learning on Mobiles

Master studies, Spring 2021/2022

Dr. Octavian Machidon
octavian.machidon@fri.uni-lj.si

University of Ljubljana
Faculty of Computer and Information Science