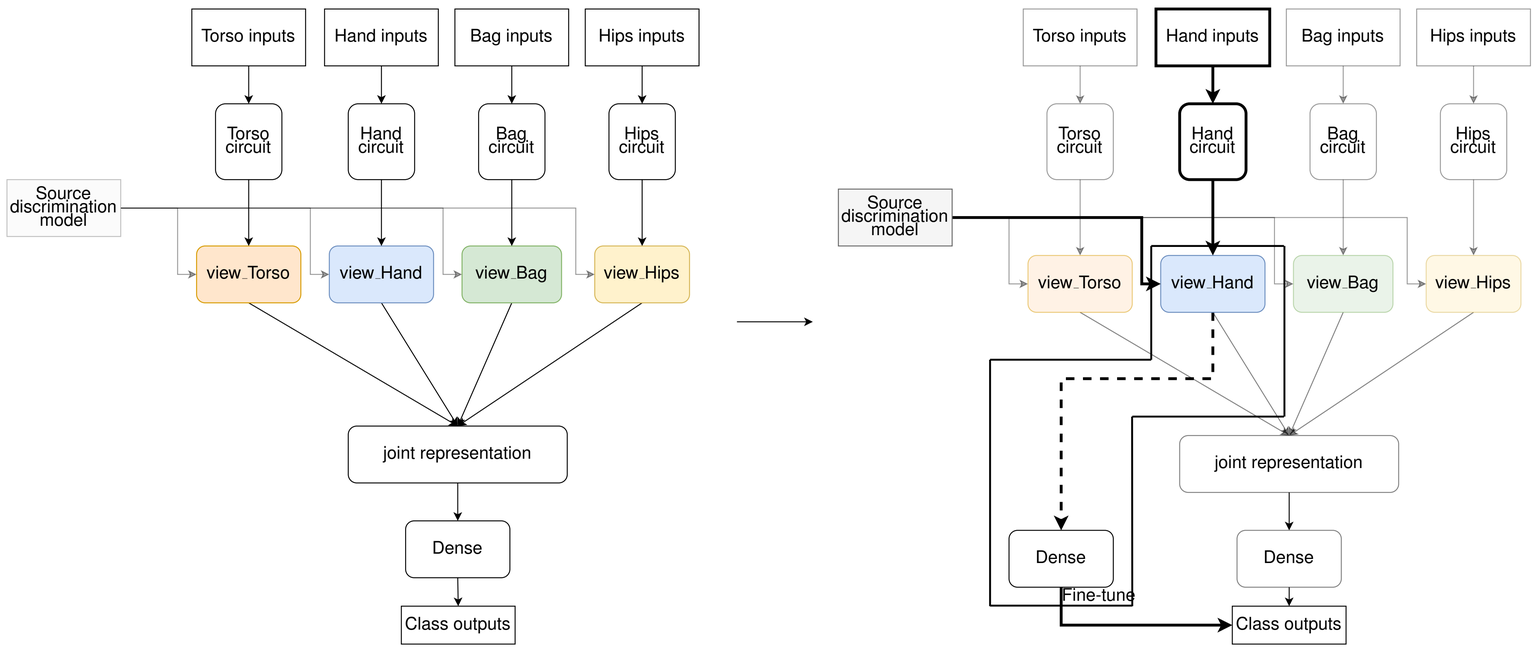
A Multi-View Architecture for the SHL Challenge

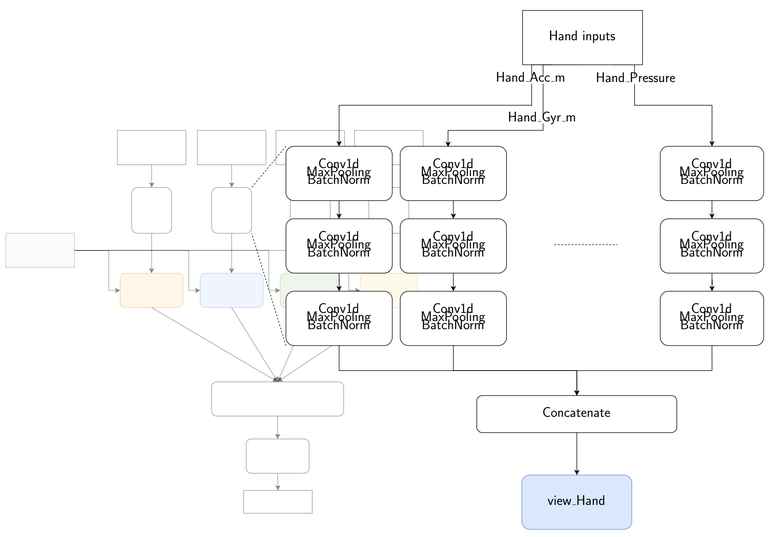
Multiple sources (positions in the case of the SHL challenge) have different levels of informativeness with regard to the concept (the transportation mode) that we want to learn.

Learning a joint-representation, as a first step, helps the model compensate for the potential lack of informativeness of some sources (e.g. a unique target position remains during model deployment, goal of this year's challenge).

Global Architecture



(1) Joint-Representation



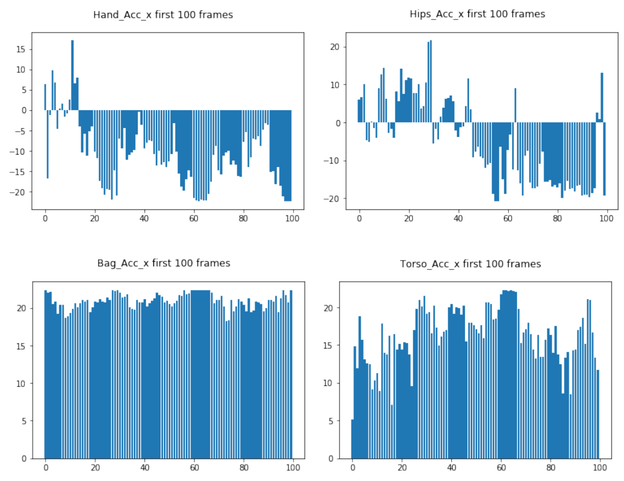
- Each phone position architecture is constructed by stacking up to 3 Conv1d/ReLU/MaxPool/BatchNorm blocks for processing each input channel individually;
- These blocks are followed by a Concatenate layer and a Dense layer (referred to view_Hand for the phone positioned on Hand).
- In this work, we use two forms of sensor channels, the **raw data** and the **magnitudes** for channels with three axes x, y and z which are calculated from the following formula and then normalized as well: $m_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$;

Using computed magnitudes rather than raw inputs improves recognition performances by more than 10%.

(2) Source Discrimination

The source discrimination model is based on the energy of input signals. We hypothesize that, while performing activities, different positions carry different energies. This is due to the fact that the amplitude of movements varies from one position to another (e.g. Hand vs. Torso).

Assume s is a signal modality of a given position with length N. The signal energy is computed as $E = \sum_{i=0}^{N-1} s_i$ where s_i is the ith sample of the s signal.

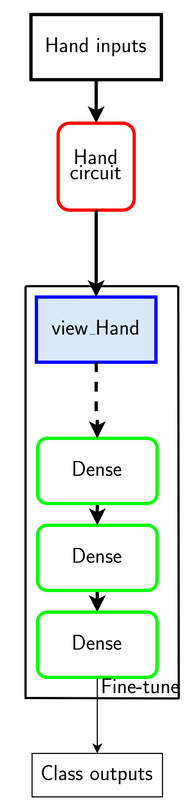


This figure shows the signal energy computed for the first 100 frames of different positions. Note that we compute these statistics for the validation dataset.

Computed signal energies of **Hand** and **Hips**, provided in the validation set, (top figures) show a similar profile as the test data.

(3) Fine-Tuning

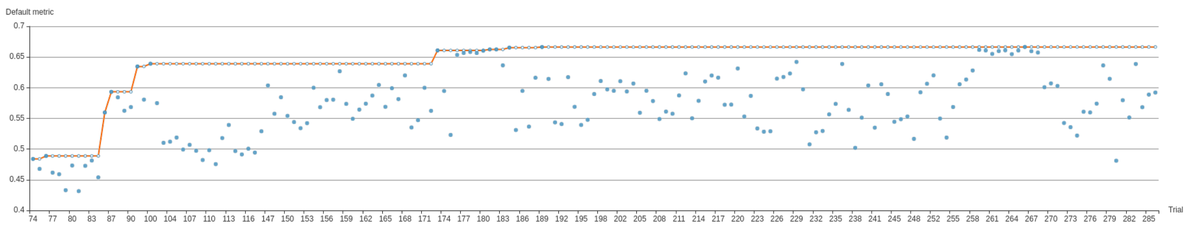
- for each source, we construct a new network using the target circuit (selected by the source discrimination model) and up to 3 additional dense layers, which are added on top of the corresponding view;
- The additional dense layers of the new network are first trained while the base network is frozen (set to inference mode). Afterward, the weights of the whole new network are trained to optimize the recognition performances;
- Fine-tuning for a specific position may produce what is referred to as "catastrophic forgetting". To compensate for the simplicity of the discrimination model, we chose to fine-tune for the target position and make sure that the final model **performs equally well** for all sources and alleviate catastrophic forgetting.



To compensate for the simplicity of the discrimination model, we fine-tune for a target position and make sure that we **alleviate** catastrophic forgetting for the remaining positions. This allows us to **handle equally-well** all sources.

Hyperparameters tuning

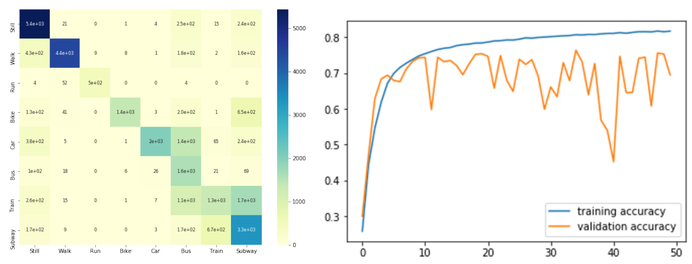
- The hyperparameters of the proposed architectures are tuned using the Tree-structured Parzen Estimator (TPE);
- TPE is a sequential model-based optimization approach which, sequentially, constructs models to approximate the performance of hyperparameters using previously explored configurations. These models are used to predict which hyperparameters instantiation to explore next;
- We use the Microsoft-NNI toolkit (`https://github.com/microsoft/nni`) which provides a comprehensive list of exploration strategies particularly based on hyperparameter tuning.



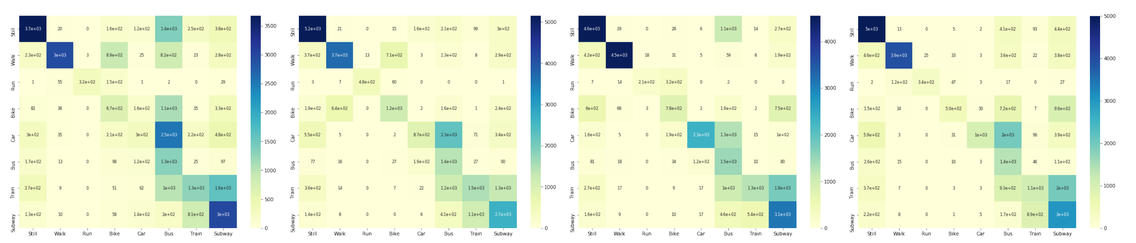
Hyperparameters tuning allows us to substantially improve the recognition performances. Noticeably, we get more than 20% improvement.

Validation on all positions

- Recognition performances of our proposed network trained on magnitude channels inputs to construct a joint representation from the different inputs;
- This network achieves approximately 81% and 75% accuracy on the training and validation sets.



Fine-tuning the whole network (valid. on Hand, Hips, Bag, Torso)



Fine-tuning individual circuits (valid. on Hand, Hips, Bag, Torso)

Fine tuning using Hips inputs **substantially improves** recognition performances on both Hips and Bag, but Hand inputs are not handled well. In contrast, when we fine-tune the model using Hand, we obtain a model that **performs equally-well on each individual position**.

Massinissa HAMIDI[1] and Aomar OSMANI[1] and Pegah ALIZADEH[2]

[1] LIPN-UMR CNRS 7030, Univ. Sorbonne Paris Nord

[2] DeVinci Research Center, Pôle Universitaire Léonard DeVinci