Machine Learning for IoT - Report Homework 1

Giuseppe Acquaviva, Mario Capobianco, Marco Mungai Coppolino
Politecnico di Torino, Italy

I. Exercise 1 - Timeseries Processing for Memory Optimization

A. Raw Data Memory Calculation

Each sensor records temperature and humidity every **2** seconds, producing $(60s \times 60m \times 24h \times 30d) \,/\, \mathbf{2} = 1,296,000$ readings per time series over the retention period of 30 days. With two time series, this results in $2 \times 1,296,000 = 2,592,000$ readings per client. Each reading requires 16 bytes (given by 8 bytes of timestamp and 8 bytes of value), for a total of $2,592,000 \times 16\,\text{bytes} = 41,472,000$ bytes ($\sim 39.55\,\text{MB}$) per client.

B. Aggregated Data Memory Calculation

Hourly aggregations (min, max, avg) for temperature and humidity generate $2 \times 3 \times (24h \times 365d) = 52,560$ records per year. The memory usage is $52,560 \times 16\,\text{bytes} = 840,960$ bytes ($\sim 0.8\,\text{MB}$) per client.

C. Total Memory Before And After Compression

The total memory per client, combining raw and aggregated data, is $39.55{+}0.8 = 40.35\,\text{MB}$. For 1000 clients, this amounts to $1000 \times 40.35 = 40,350\,\text{MB}$ ($\sim 39.41\,\text{GB}$). Assuming 90% compression, the total memory required to provide the monitoring service becomes $39.40 \times 0.1 = 3.94\,\text{GB}$

D. Discussion

Raw data constitutes 98% of memory usage, with aggregated data requiring significantly less due to hourly sampling. Compression reduces total memory requirements to $3.94\,\text{GB}$

II. Exercise 2 - Voice Activity Detection Optimization and Deployment

A. Methodology

Our approach to determine optimal VAD hyperparameters systematically explores configurations aligned with specified constraints in audio processing, including sampling rate, frame length, frame step, dB threshold and duration threshold. After deciding suitable parameter ranges, we performed a grid search, varying the frame length $= l$ (using powers-of-2 and non-powers, both in the $[10, 50]$ ms range), the frame step $= s$ (by changing the overlap $\{0\%, 25\%, 50\%, 75\%\}$, with the formula: $s = l \times (1 - \text{overlap})$), the dB threshold (5, 10, 20 and 40 dB), and the duration threshold (0.1s to 0.5s).

B. Table and Comments

The grid search results are summarized in Table I.

TABLE I
VAD Hyperparameters and Results Comparison

Desc.	SR	FL(s)	FS(s)	dBthres	Dur_thres(s)	Acc (%)	Lat. (ms)
Baseline	16000	0.04	0.01	20	0.4	65.11	47.1
Ours	16000	0.032	0.016	10	0.15	97.89	19.9

C. Discussion of Table Results

a) Frame Length (FL): `frame length` affects both **latency** and **accuracy**. Regarding latency, it was hard to predict in advance the effect of the hyperparameter. While longer lengths reduce the total number of frames to process, they increase the computational cost of performing the FFT on each frame. Using frame lengths, such as 16 ms or 32 ms, that result in power-of-2 sample sizes improves the efficiency of the FFT and enhances the quality of the results. Among these, a frame length of 32 ms reached a favorable balance, providing low latency, while maintaining high accuracy.

b) Frame Step (FS): a smaller `frame step` increases the **latency** (this is due to a higher number of frames) but produces a better **accuracy**, vice-versa increasing s reduce the computation time, worsening the classification. From our tests we saw that an overlap of 50% was preferred to achieve a better compromise between speed and precision.

c) dB Threshold (dBthres): the `dBthres` hyperparameter strongly influences **accuracy** without affecting latency. Thresholds of 20 and 40 dB were insufficiently sensitive to energy differences between speech and silence, while 5 dB was overly sensitive to noise. A threshold of 10 dB provided a balanced trade-off between noise robustness and speech detection, making it the optimal choice in most tests.

d) Duration Threshold (Dur_thres): the `duration_thres` impacts **accuracy**. Changing this value does not impact latency. Excessively short thresholds may increase false positives by misclassifying transient noises as speech. The chosen threshold (0.15s) effectively balances the need for rapid detection while minimizing errors. If the duration threshold is too high (e.g., 0.5s), the VAD may have failed to detect shorter speech segment.

D. Matching Target Constraints

The values reported in Table I obtained an accuracy $>$ 97.6 and a latency $<$ 25 ms, meeting the target constraints. By empirically selecting the optimal frame length from the computationally efficient options and by reducing the frame overlap, we achieved a faster processing compared to the one of the default settings.