Collaborative Visual Place Recognition through Federated Learning

Mattia Dutto*, Gabriele Berton, Debora Caldarola, Eros Fanì, Gabriele Trivigno, Carlo Masone
Politecnico di Torino
name.surname@polito.it
*Corresponding Author: Mattia Dutto

Appendix

The Appendix is organized as follows:
- Appendix A: additional details on centralized runs.
- Appendix B: implementation details.
- Appendix C: additional analyses on MSLS Proximity split.
- Appendix D: ablation studies on Hierarchical Federated Learning.

A. Centralized runs

This section introduces additional details on the centralized experiments presented in **??**. Centralized training continues until the network performance plateaus for five consecutive epochs. This approach results in training different networks for a varying number of epochs depending on their convergence speed. Tab. 1 compares the selected model architectures in terms of number of epochs, recall and training time. Based on these results and on the number of parameters (**??**), we select ResNet18 truncated as our network.

Table 1. **Centralized baselines**: model architectures compared in terms of number of epochs, recall (R@1) and training time in the centralized scenario.

Backbone	Epochs	R@1	Time
ResNet18 truncated	40	42.9 $\pm$ 2.5	15h30
ResNet18	31	60.1 $\pm$ 0.3	10h15
VGG16	30	46.3 $\pm$ 0.5	28h30

B. Implementation details

This section extends **??** with additional implementation details on our experiments. The codebase is written in Python with PyTorch for neural networks optimization. The experiments are run on the Nvidia Titan X GPU with 12GB of VRAM. All runs are averaged across 3 different seeds.

Model. The experiments are run using a ResNet18 truncated after the third convolutional layer. The pooling layer is GeM except for the baseline experiments available in **??** where we tested SPOC and MAC as well. The image resolution is always 288x384 pixels except for the cases in **??** and **??** where we tried different values: 96x128, 192x256, 384x512, and 480x640 pixels.

Table 2. Server-side optimizers hyperparameters (learning rate η_s and momentum β_s).

Method	η_s	β_s
FedAvg	1	0
FedSGD	0.1	0.9
FedAdam	0.1	0.9
FedAdaGrad	0.01	0.9

Table 3. Number of rounds T and selected clients per round C when comparing FedAvg with H-FL.

Method	T	C
FedAvg	75	20
H-FL Continent	75	20 (4 continents by 5 clients per continent)
FedAvg	15	105
H-FL City	15	105 (21 cities by 5 clients per city)

FL baselines. The number of rounds T in the FL experiments is set to 300, with 5 clients per round. The server optimizer is always SGD with learning rate 1, *i.e.*, FedAvg. Tab. 2 reports the hyperparameters used for the ablation on the server-side optimizers (SERVEROPT in **??**) from **??**. In local training, the optimizer is Adam with learning rate is always set to $1e-5$ and momentum 0. Each client runs one epoch. Unless otherwise specified, the maximum number of local iterations is set to 2500. When comparing H-FL with FedAvg (**??**), we modify the number of rounds T and participating clients C accordingly, as summarized in Tab. 3.

Data augmentation. We study the effect of data augmentation techniques in **??**. Due to the required increased time, data augmentation is not used by default in the other experiments. In **??**, data augmentation is applied with 50% probability. We apply color jitter (hue, saturation, brightness, and contrast) and random resize crop. Normalization instead is always applied with standard ImageNet values.

Table 4. **Comparison of different aggregation interval T_s in h-FL.** The experiments are run with the continental aggregation with 5 clients per continent at each round and carried on for 300 rounds.

T_s	R@1
5	60.1 ± 0.8
10	60.4 ± 0.9
15	**61.1 ± 0.6**
20	60.0 ± 0.9
25	60.3 ± 0.4

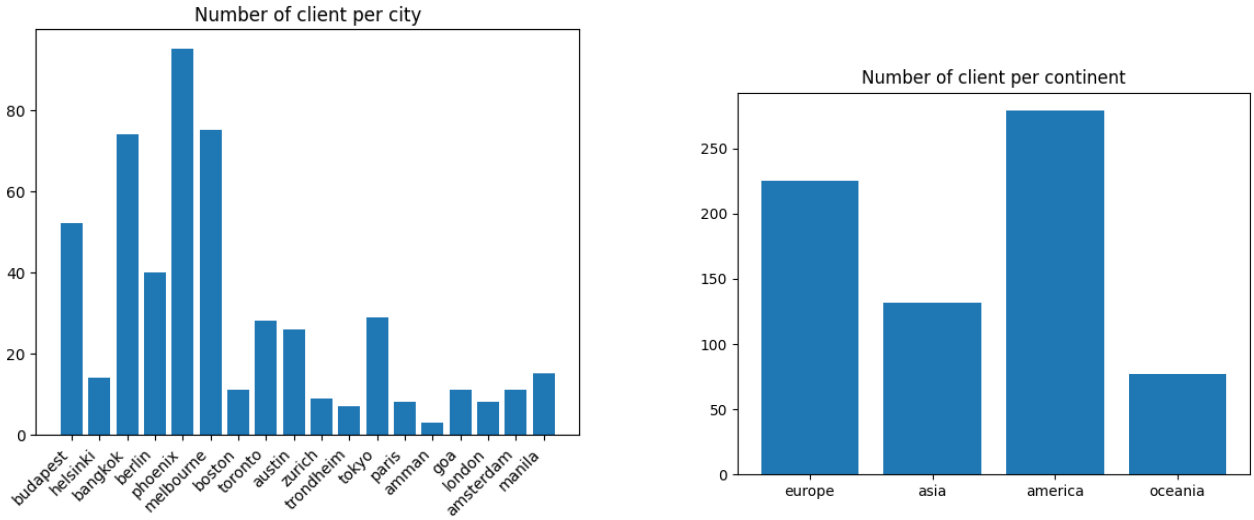
Local mining. We set the number of sequences for computing the mining dataset to 333 and 20 and the number of images selected per sequence to 3 and 50 respectively.

C. Distribution of clients in federated MSLS

In this section, we present additional analyses on the MSLS federated splits described in **??**. Focusing on the *Proximity* split, Fig. 1a shows the distribution of clients across cities. We note that Budapest, Bangkok, Phoenix and Melbourne are the most populated. Fig. 2a shows that those same cities are also the ones containing most images. Figs. 1b and 2b repeat the same analyses per continent: even if most of the clients are found in America, Europe has most of the images, while the least populated continent is Oceania but Asian clients have in total less images.

D. Ablation studies on H-FL

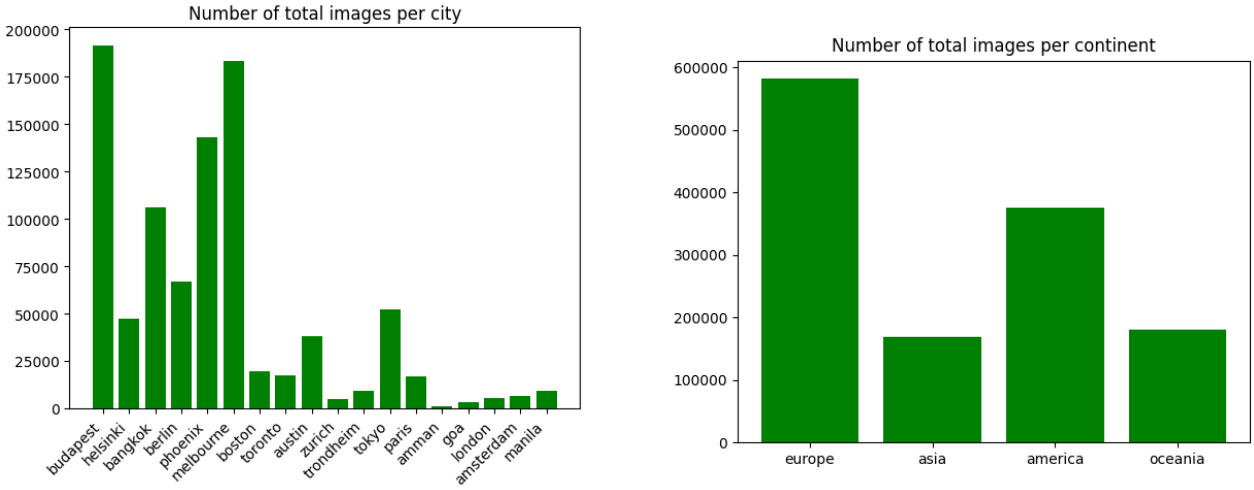
Tab. 4 investigates the effect of the round interval (T_s) between aggregation steps in H-FL for continent-based clustering. The results show that an optimal value exists for T_s. Setting T_s too low hinders the cluster-specific models from learning generalizable information, while a very high T_s leads to reliance on outdated updates. The best performance is achieved with $T_s = 15$.



(a) Distribution of clients per **city**

(b) Distribution of clients per **continent**

Figure 1. Clients distribution in the MSLS Proximity split.



(a) Distribution of images per **city**

(b) Distribution of images per **continent**

Figure 2. Images distribution in the MSLS Proximity split.