NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Semantic-aware Visual Localization With Attention

Student: Zhou Zeyu Supervisor: Lin Weisi Co-supervisor: Hu Zhi

Introduction and Problem:

Visual Place Recognition in the urban environment is very challenging, because the scene may have both dynamic objects (cars, people, and sky) and static objects (buildings, roads, and terrain). In order to let the model focus on the static objects, an attentional mechanism is generally used by assigning each feature a relevance weight. However, such approaches **cannot precisely label every feature** correctly.

Hypothesis:

The current attention method is not perfect. If a **deterministic binary segmentation filter** can be added to the original attention mechanism, the result of visual place recognition may be improved. The filter which is trained as semantic segmentation model may better distinguish static from dynamic objects.

Methodology:

1. Utilize SOTA semantic segmentation model Mask2Former to label all train set images.
2. Train a U-Net segmentation model on the labelled dataset created in the last step, to separate static and dynamic objects.
3. Generate feature map using VGG16 and obtain feature inter-weights from semantic-initialized NetVLAD layer.
4. Apply the binary filter to the weighted feature map from last step as shown in figure 1.
5. Aggregate the feature and obtain the final image descriptor as VLAD.

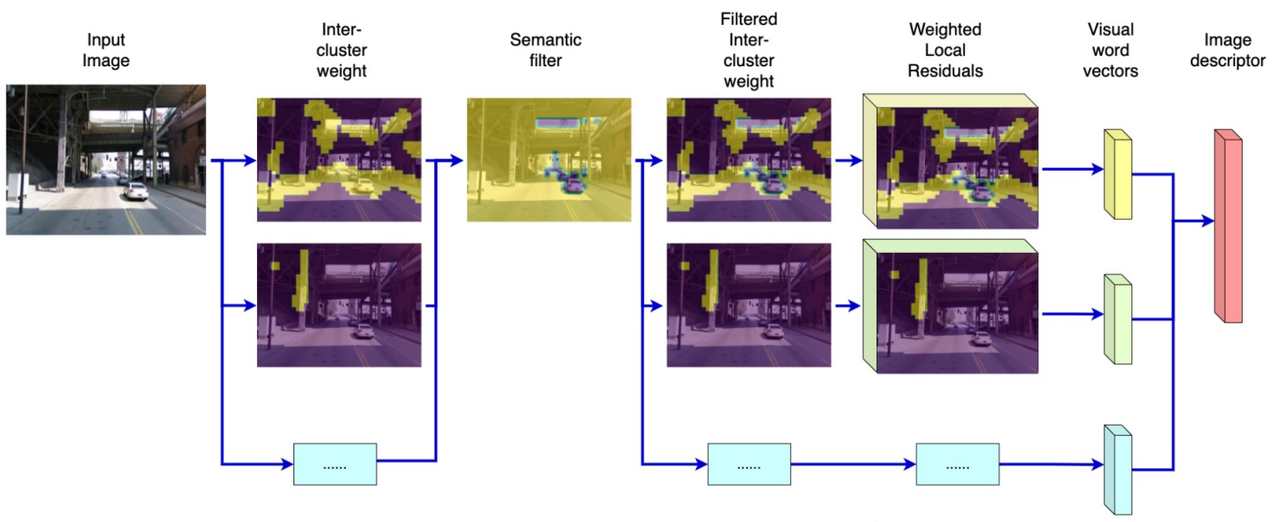

Figure 1: How to apply the semantic filter

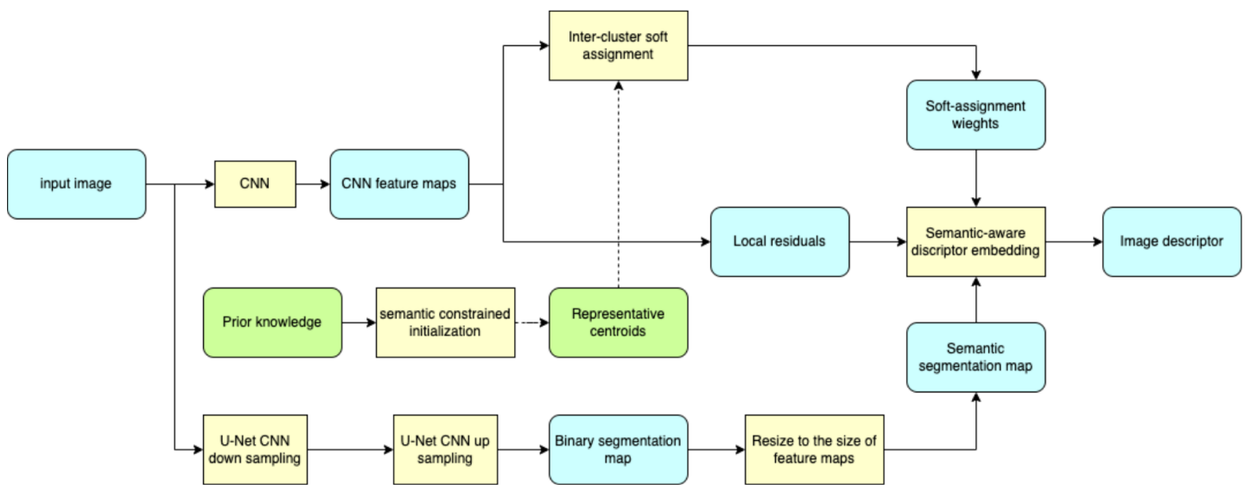

Figure 2: Overall structure

Experiment:

The results from 4 different models (**NetVLAD, SRALNet,U-Net filter, and Mask2Former**) are compared. SRALNet, U-Net Filter, and Mask2Former are sharing with the same semantic initialization data.
From the figure 3, we can observe that the Mask2Former can perfectly distinguish static from dynamic features. U-Net can also perform relatively well on segmentation of the image.
From the numeric result, Mask2Former filter achieves the highest recall rate among all, U-Net Filter can also achieve a similar recall as the Mask2Former.

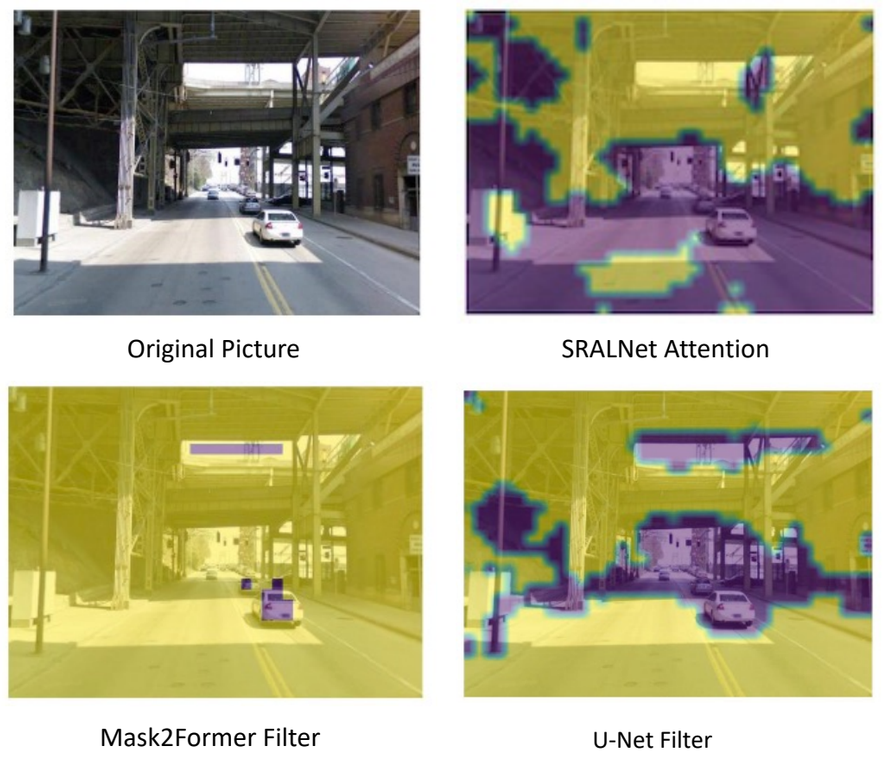

Figure 3: Compare between attention maps

Model	Recall@1	Recall@5	Recall@10
NetVLAD	83.6	92.1	93.6
SRALNet	84.5	93.9	96.4
U-Net Filter	85.7	94.0	97.5
Mask2Former Filter	86.0	95.0	98.3

Conclusion and Future Work:

The result has **clearly indicated** that **the deterministic binary filter** is able to improve the performance of the visual place recognition model.
However, currently, the U-Net model based on VGG16 features cannot segment the image perfectly. In the future a deeper backbone like Resnet50 should be used to generate the feature map.
Besides, current filter is binary-based, assigning the same weight to all static features. However, it is obvious that different static objects may have different importance to the visual place recognition task. Thus, the filter may be improved by assigning different weights to different categories of objects.