Dual-Gradients Localization framework for Weakly Supervised Object Localization

Chuangchuang Tan [1*], Tao Ruan [1*], Guanghua Gu [2],
Shikui Wei [1], Yao Zhao [1]

[1] Beijing Jiaotong University

[2] Yanshan University

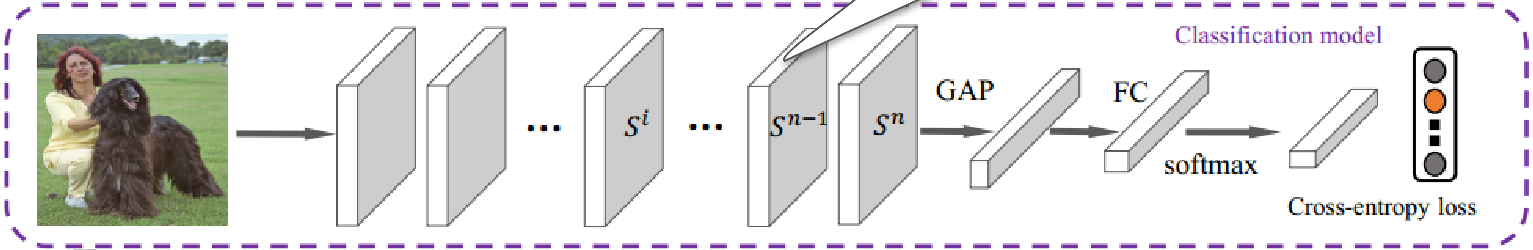
Target

- Weakly Supervised Object Localization (WSOL)
  - WSOL is understanding an image at pixel level only using image-level annotations
  - use much cheaper annotations

WSOL

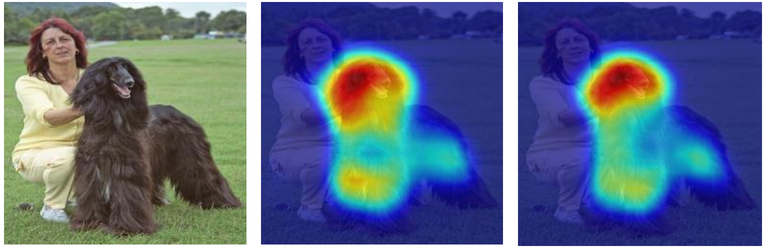
- Steps of previous works：
  - Force classification network to focus on more regions of feature map.
  - Produce localization map on the last convolutional layer by applying CAM.
- Problem:
  - ignore the localization ability of other layers.
  - Both localization and classification tasks are trained online

I can produce WSOL, too



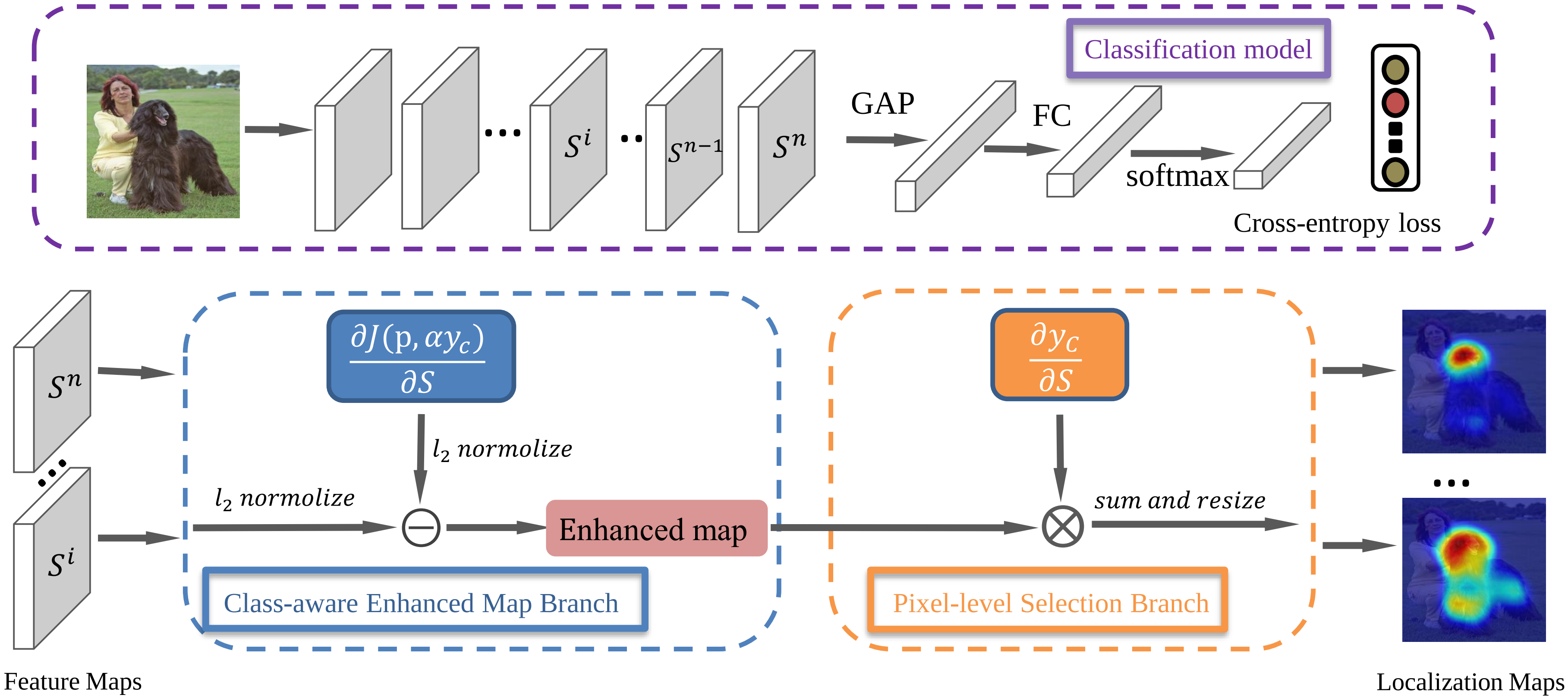
Dual-Gradients Localization(DGL) framework

- Main ideas：
  - Utilize gradients of classification loss function to mine entire target object regions.
  - Leverage gradients of target class to identify the correlation ratio of pixels to the target class within any convolutional feature maps
- Characteristics
  - Simple, DGL is a offline approach, needn’t to train for localization.
  - Effective, achieving localization on any convolutional layer.

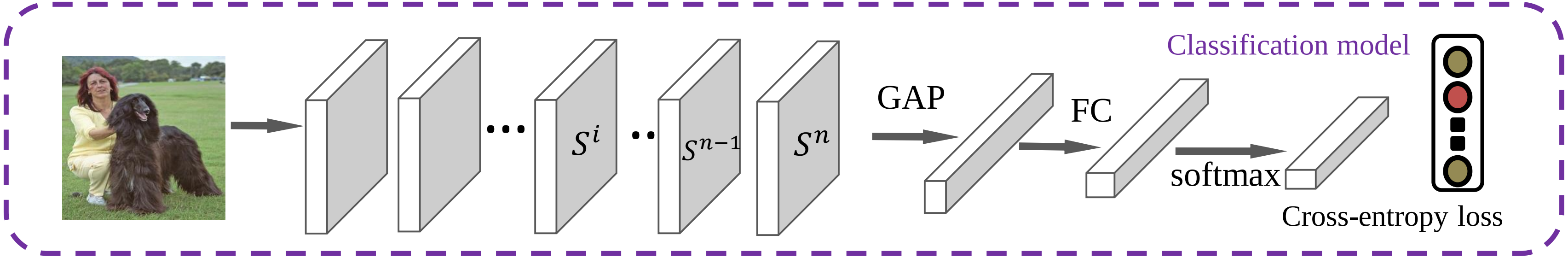


Source image Mixed_6f Mixed_6e

Overview of the DGL framework



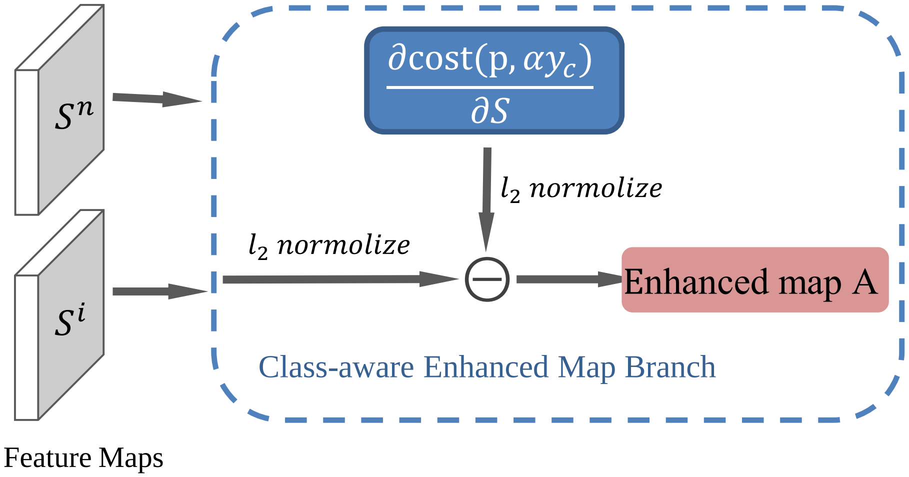
Classification model

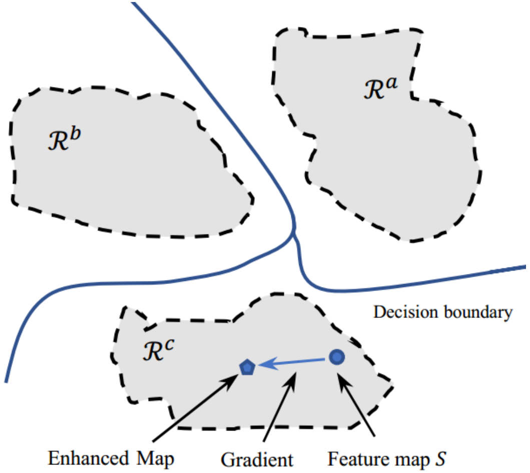


- Classification model architecture：
  - use a customized InceptionV3, i.e. SPG-plain.
  - remove the layers after the second Inception block, i.e., the third Inception block, pooling and linear layer.
  - add two convolutional layers
  - add a GAP layer and a softmax layer

Class-aware Enhanced Map Branch

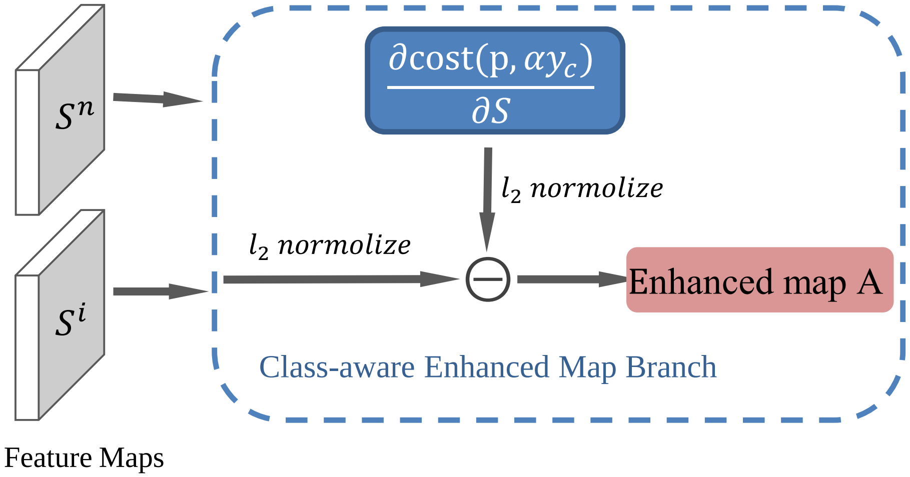
- feature maps predicted to class c only capture the discrimination parts of objects, when the feature maps close the boundary of classification regions
- the feature maps located at center of classification regions can highlight more object regions

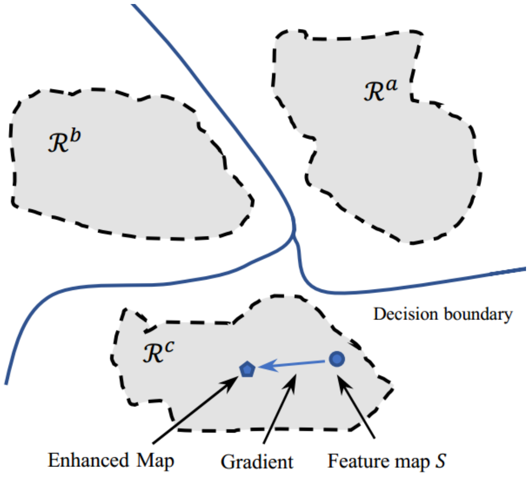




Class-aware Enhanced Map Branch

- our key idea of Class-aware Enhanced Map is pulling the feature maps toward inside of the classification region for specific-class, along with gradients of classification loss function.

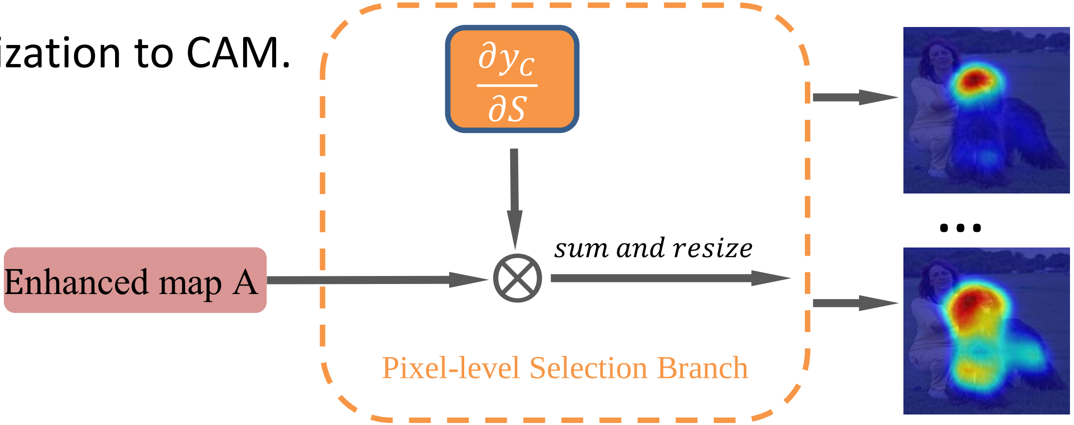




- Is gradients or weights?
  - CAM actually achieves localization by employing a weighted sum of feature maps and gradients of target class on the last convolutional layer, instead of weights of the final FC layer.

  - Pixel-level Selection is a generalization to CAM.

$$\overline{M}_c^i = \sum_k \frac{\partial l_c}{\partial S_k^i} \{A_c^i\}_k$$



Results on the Validation Set of LID

MS: Multi-scale inputs during test
MC: Morph close the localization map during test

MS	MC	mIoU
✘	✘	58.23
✔	✘	61.46
✔	✔	62.22

- Fusion the localization maps of branch1 and branch2 on Mixed_6e layer.
- Input size 324

Qualitative Results

- Examples of DGL on test set

北京交通大學
Beijing jiaotong University

燕山大學
YANSHAN UNIVERSITY