High-Resolution Class Activation Mapping
Thanos Tagaris*, Maria Sdraka and Andreas Stafylopatis
*thanos@islab.ntua.gr
National Technical University of Athens

Code Available at github.com/djib2011/high-res-mapping

### Abstract
- Neural Networks can’t provide sufficient reasoning for their decisions. They operate like black boxes. This is especially true in Deep Neural Networks.
- What a Convolutional Neural Network (CNN) could also do is measure responsibility of regions in the input image by production of Class Activation Maps (CAMs).
- Compared to related work, our approach generates robust, refined and low-resolution Class Activation Maps, without impairing the CNN’s performance.

### Proposed Localization Framework
The proposed framework consists of 3 parts: the Localization Network, the Expansion Network and the Postprocessing Pipeline. Compared to similar frameworks, it is capable of:
1. Producing high-resolution, refined and robust Class Activation Maps.
2. Maintaining the state-of-the-art performance.

### Localization Network
This is a regular CNN with the ability of producing CAMs.

### Expansion Network
The “Expansion Network” is built to mirror the Localization Network.

### Combined Model Architecture

<table>
<thead>
<tr>
<th>Classification / localization network</th>
</tr>
</thead>
<tbody>
<tr>
<td>224 x 224 x 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expansive path</th>
</tr>
</thead>
<tbody>
<tr>
<td>224 x 224 x 3</td>
</tr>
</tbody>
</table>

### Experimental Procedure
- The framework was evaluated on the animals classes from the ImageNet dataset. This offers the unique characteristic of having a large number of classes (i.e. 398) that are very similar to one another (e.g. ‘tabby cat’, ‘tiger cat’, ‘Persian cat’, ‘Siamese cat’, ‘Egyptian cat’)
- The first step is to pre-train the Localization Network for classification.
- Afterwards, the combined architecture (i.e. localization and expansion) is trained in unsupervised fashion, while keeping the localization network’s weights frozen.
- To provide a Refined CAM, the two CAMs generated from the network (i.e. low and high-res) are passed through the Postprocessing pipeline.

### Example Pipeline
Figure: Left: an example image. Right: the low-res CAM.

### Class Activation Mapping
- A Class Activation Map (CAM) is the region of the input image that the CNN uses to predict a given class.
- A CNN can produce CAMs if it concludes with a Global Average Pooling (GAP) and a Fully Connected (FC) layer.

### Motivation
- Even though Deep Neural Networks have achieved state-of-the-art performance in several image and text related tasks, they have experienced a slow adoption rate by some industries.
- More specifically, in sensitive domains where mistakes matter and there is a notion of responsibility and accountability.
- Some examples are the fields of medical diagnoses and autonomous driving. Even though research indicates that Deep Learning algorithms surpass human performance, these techniques haven’t been adopted in these fields. The issue of trust is the main reason.
- If there was a way for a CNN to provide a reasoning for its decisions, then it would be easier for humans to trust it.

### Postprocessing Pipeline
The goal of this pipeline is to combine the low and high-resolution maps into a single Refined CAM. To accomplish this the following postprocessing steps were drafted:
1. The Low-Resolution CAM contributes to the Refined CAM in two ways: Its focal points are identified and a Region of Interest (ROI) is extracted.
2. The high-resolution CAM is first blurred then passed through a Sobel filter for edge detection.
3. A threshold-based Region Growing segmentation technique is applied on the previous high-res map, using the data (i.e. starting seeds, threshold values) from the low-res one.
4. The resulting segmented map is combined with the upscaled low-res one, after having having filled its small “holes”, to produce the Refined CAM.

### Proposed Architecture

![Combined Model Architecture Diagram]

---

DenseNet

- A network capable of achieving a top-5 error rate of 3.6% on the ImageNet dataset.

- Its goal is to take the low-resolution feature maps and expand them to the dimensions of the original image, used to produce the High-Resolution CAMs.

- The architecture elected is DenseNet, a network capable of achieving a low-res CAM.

- In order to maintain state-of-the-art performance, the network needs to have low-resolution feature maps before its GAP layer.

- The architecture includes both the Localization and Expansion Networks.

---

**Motivation**
- Even though Deep Neural Networks have achieved state-of-the-art performance in several image and text related tasks, they have experienced a slow adoption rate by some industries.
- More specifically, in sensitive domains where mistakes matter and there is a notion of responsibility and accountability.
- Some examples are the fields of medical diagnoses and autonomous driving. Even though research indicates that Deep Learning algorithms surpass human performance, these techniques haven’t been adopted in these fields. The issue of trust is the main reason.
- If there was a way for a CNN to provide a reasoning for its decisions, then it would be easier for humans to trust it.

**Class Activation Mapping**
- A Class Activation Map (CAM) is the region of the input image that the CNN uses to predict a given class.
- A CNN can produce CAMs if it concludes with a Global Average Pooling (GAP) and a Fully Connected (FC) layer.

**Motivation**
- Even though Deep Neural Networks have achieved state-of-the-art performance in several image and text related tasks, they have experienced a slow adoption rate by some industries.
- More specifically, in sensitive domains where mistakes matter and there is a notion of responsibility and accountability.
- Some examples are the fields of medical diagnoses and autonomous driving. Even though research indicates that Deep Learning algorithms surpass human performance, these techniques haven’t been adopted in these fields. The issue of trust is the main reason.
- If there was a way for a CNN to provide a reasoning for its decisions, then it would be easier for humans to trust it.