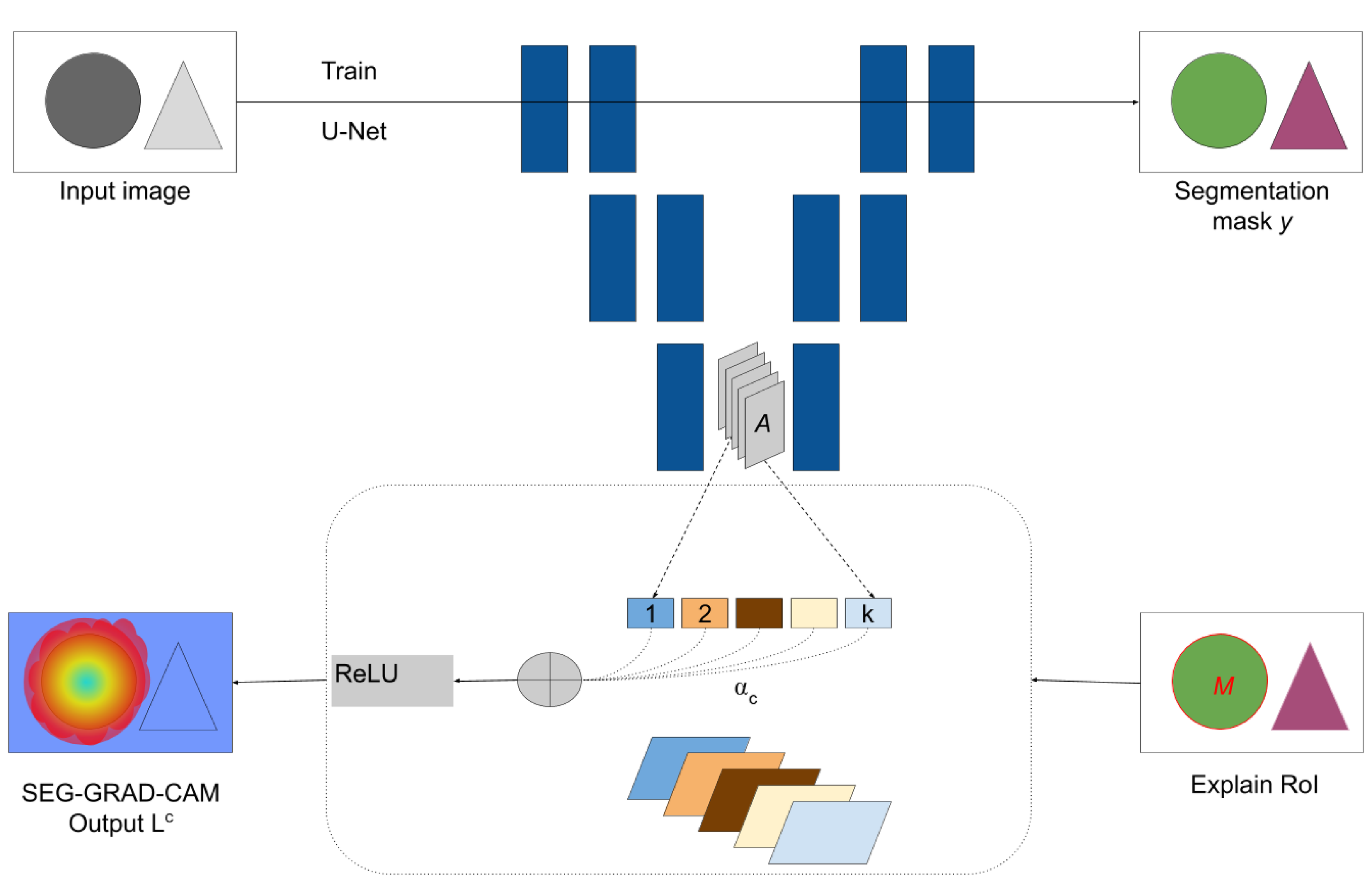
Towards Interpretable Semantic Segmentation via Gradient-weighted Class Activation Mapping

Kira Vinogradova (vinograd@mpi-cbg.de), Alexandr Dibrov, Gene Myers
Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
Center for Systems Biology Dresden, Germany

Introduction

- Semantic segmentation is an important common task
- Explainability / Interpretability is an active area
- Various methods exist for interpretation of classification networks
- This is one of the first approaches to explain semantic segmentation
- SEG-GRAD-CAM = segmentation Grad-CAM [1]
- SEG-GRAD-CAM is applied locally to produce heatmaps showing the relevance of a set of pixels or an individual pixel for semantic segmentation.

Method



Grad-CAM on a set of pixels of the predicted mask

SEG-GRAD-CAM is based on Grad-CAM [1].

Grad-CAM averages the gradients of the logit y^c of class c with respect to all N pixels (indexed by u,v) of each feature map A^k to produce a weight α_c^k to denote its importance. $\alpha_c^k = \frac{1}{N}\sum_{u,v}\frac{\partial y^c}{\partial A_{uv}^k}$ (1)

In Grad-CAM: A^k are taken from the last convolutional layer of a classification network, in SEG-GRAD-CAM: from a bottleneck layer of U-Net [2].

The heatmap L_c is the weighted non-negative sum of the feature maps. $L_c = ReLU(\sum_k \alpha_c^k A^k)$ (2)

A CNN for semantic segmentation typically produces logits y_{ij}^c for every pixel x_{ij} and class c. In SEG-GRAD-CAM, y^c is replaced by $\sum_{(i,j)\in M} y_{ij}^c$ where M is a set of pixel indices of interest in the output mask.

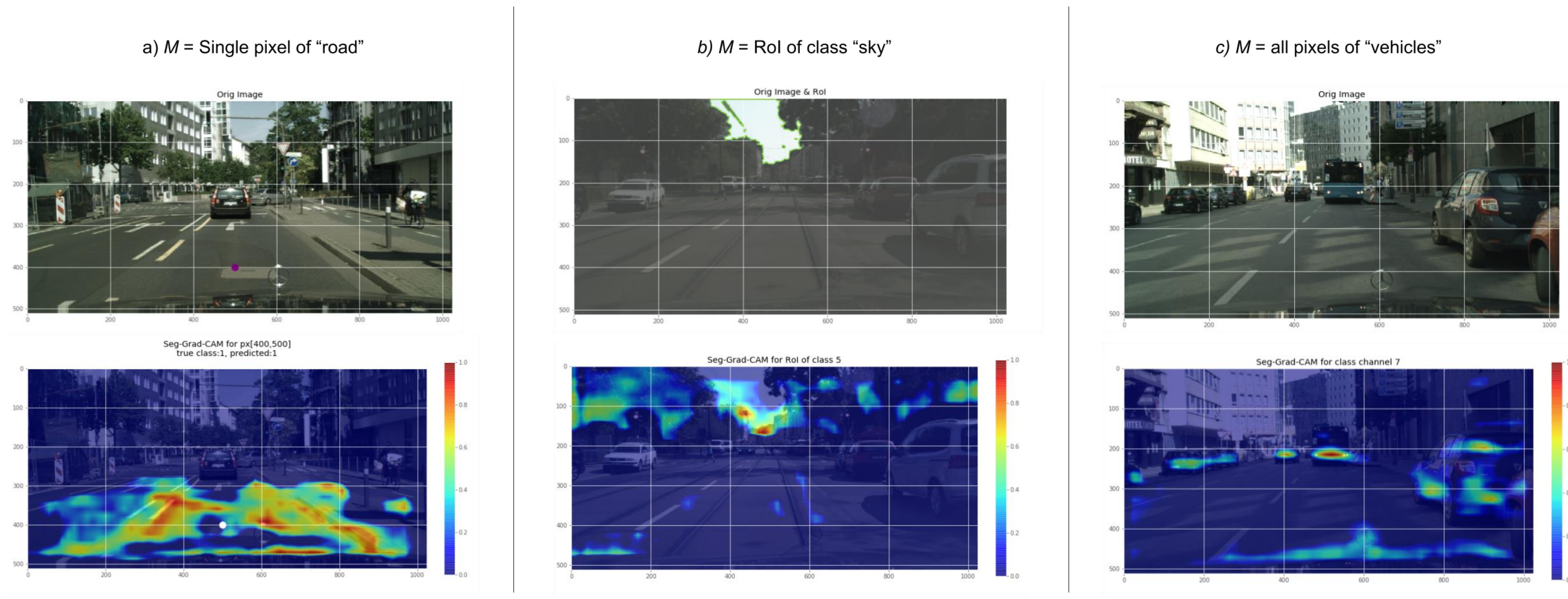
The region of interest can be a single pixel, or an object instance, or all pixel classified as c.

Formula of SEG-GRAD-CAM saliency map:

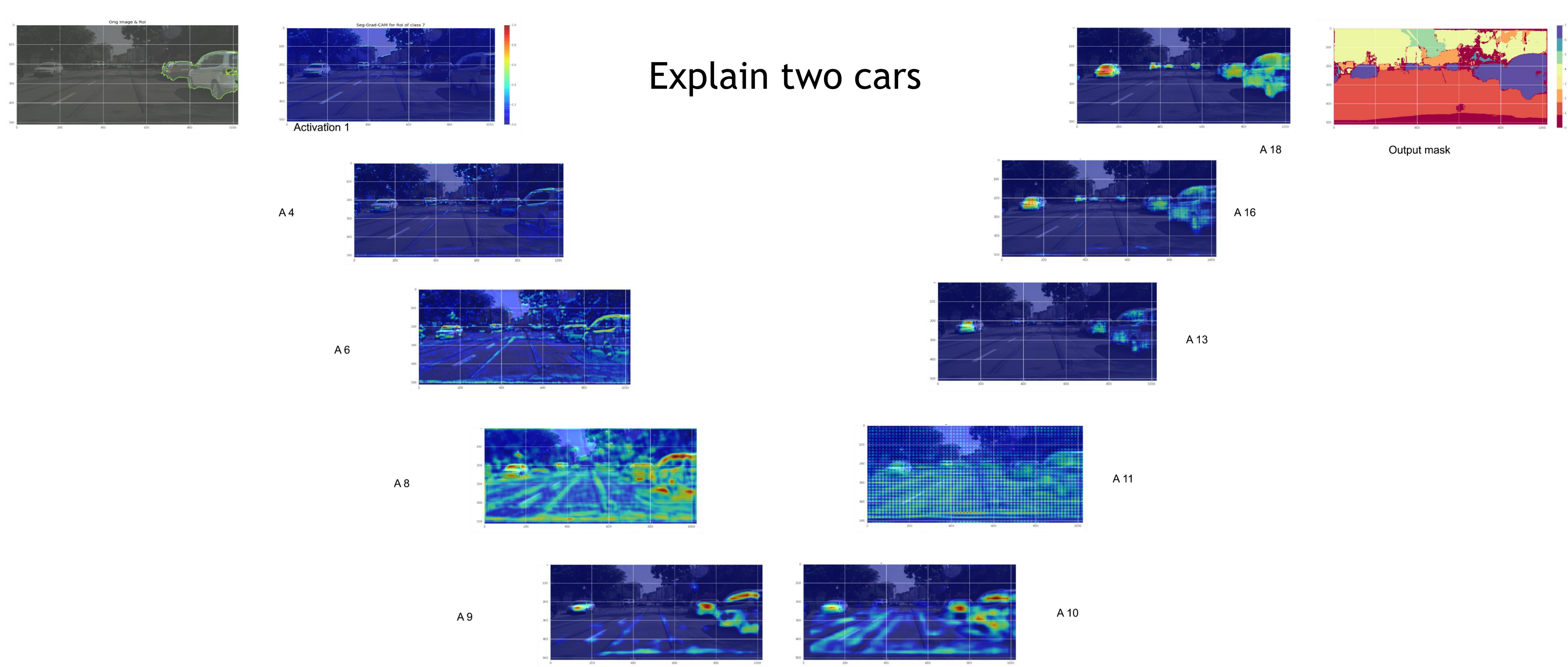
$$L_c = ReLU\left(\sum_k A^k \frac{1}{N}\sum_{u,v}\frac{\partial(\sum_{(i,j)\in M} y_{ij}^c)}{\partial A_{uv}^k}\right) \quad (3)$$

Presented at AAAI in Feb 2020

Results



SEG-GRAD-CAM can produce saliency maps for any subset of pixels. *a)* shows relevance for a single pixel. *b)* demonstrates relevance for a region of interested, e.g. one of the cars, or the hood of the Mercedes, or a piece of sky. *c)* shows the case in which M is a class channel c in the predicted mask in formula (3). We trained a U-Net [2] architecture on *Cityscapes* [3] on 8 categories: void, flat, construction, object, sky, human, vehicle.



The above figure aims to explain the choice of the bottleneck layers as the suitable layers to retrieve feature maps A. The heatmaps produced from the initial convolutional layers exhibit edge-like structures. Feature maps from the bottleneck demonstrate aspects of the object and the context. Intuitively, the bottleneck contains a condensed representation of objects' characteristics. Feature maps located further look more and more similar to the logits of the selected class and the output mask.

References

[1] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In ICCV.

[2] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In MICCAI.

[3] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In CVPR.