Visual Feature Attribution using Wasserstein GANs

Christian F. Baumgartner[1] Lisa M. Koch[2] Kerem Can Tezcan[1] Jia Xi Ang[1]
Ender Konukoglu[1] for the Alzheimer's Disease Neuroimaging Initiative*

[1]Computer Vision Lab, ETH Zurich [2]Computer Vision and Geometry Group, ETH Zurich

CVPR 2018

Presenter: Jack Lanchantin

Visual Attribution Methods

- Most visual attribution methods training a classifier to predict the class and then use one of the following:
  - Saliency maps (gradient of class w.r.t image)
  - Activation maps (activations of the feature maps during classification)

Visual Attribution Methods

- Shwartz-Ziv & Tishby showed that during training, NNs minimize the mutual information between input and output layers, thus compressing input features
    - The model may ignore features with low discriminative power if stronger features are available.
    - If there is evidence for a class at multiple locations in the image some locations may not influence the classification and may not be detected
  - ➢ Training may be working in opposition to the goal of visual attribution

This paper

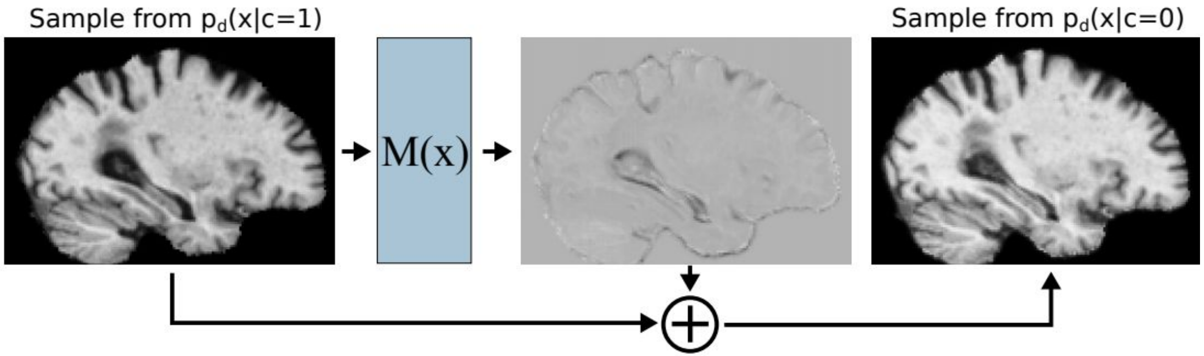
- Try to visualize evidence of a particular category in a way that captures all category-specific effects in an image.
- Find a map s.t. when added to image of one class, changes to another class
- 2 Differences between previous methods:
    - Does not rely on a classifier (assumes test image categories have already been determined)
    - Requires a baseline class (e.g. benign MRI image)

Problem Formulation

- Given:
  - Classes $c \in \{0, 1\}$, a baseline class and a class of interest
  - Image x
  - Distribution of images from class $c = 0$ with $p(x|c = 0)$
  - Distribution of images from class $c = 1$ with $p(x|c = 1)$

Problem Formulation

Estimate a map function $M(\cdot)$ that, when added to an image x_i from category $c = 1$, creates an image $y_i = x_i + M(x_i)$ which is indistinguishable from the images sampled from $p(x|c = 0)$.



Visual Attribution GAN (VAGAN)

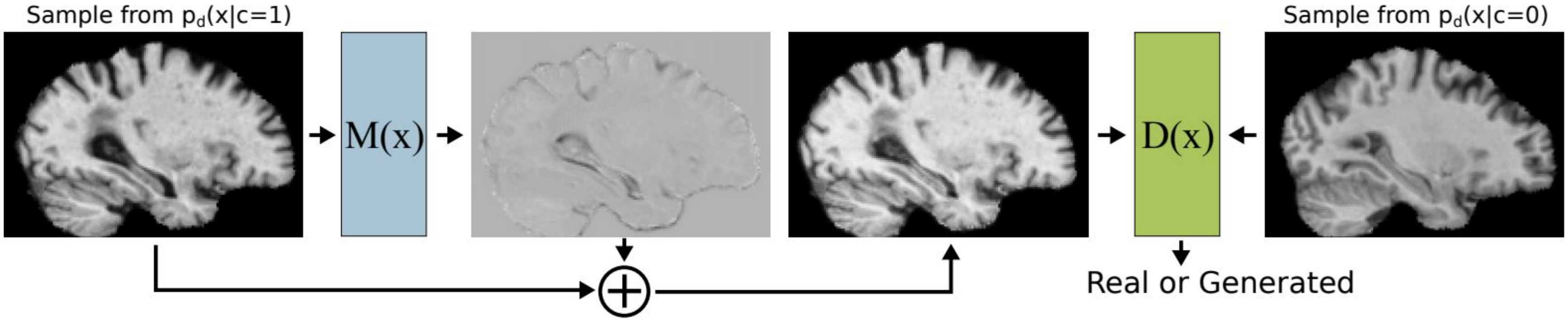


Figure 2. Overview of VA-GAN. During training images are sampled from the categories $c \in \{0, 1\}$. Images from $c = 1$ are passed to the map generating function $M(x)$. The map generator aims to create additive maps which produce generated images that the critic $D(x)$ cannot distinguish from images sampled from $p_d(x|c = 0)$. The critic, $D(x)$ tries to assign different values to generated and real images. During testing, $M(x)$ can be used directly to predict a map in a single forward pass.

Visual Attribution GAN (VAGAN)

$$\mathcal{L}_{GAN}(M, D) = \mathbb{E}_{x\sim p_d(x|c=0)}[D(x)] - \mathbb{E}_{x\sim p_d(x|c=1)}[D(x + M(x))].$$

$$\mathcal{L}_{reg}(M) = ||M(x)||_1$$

$$M^* = \underset{M}{\operatorname{argmin}} \max_{D\in\mathcal{D}} \mathcal{L}_{GAN}(M, D) + \lambda\mathcal{L}_{reg}(M)$$

where *D* is the set of 1-Lipschitz functions

Baseline Approach - Additive Perturbation Maps

- Train a classifier *f*(x) = p(c = 1) and then optimize map *m* to lower p(c = 1)
    - I.e. the image $y_i = x_i + m$ should minimize $f_i(y_i)$
    - Similar to VAGAN except that *m* is not a function of x_i
- Finding image map *m* involves minimizing:

$$m^* = \underset{m}{\operatorname{argmin}} f(x_i + m) + \omega_1 ||m||_1 + \omega_2 \sum_u ||\nabla m(u)||_\beta^\beta.$$

where *u* are the pixels of *m*

Synthetic Data Experiments

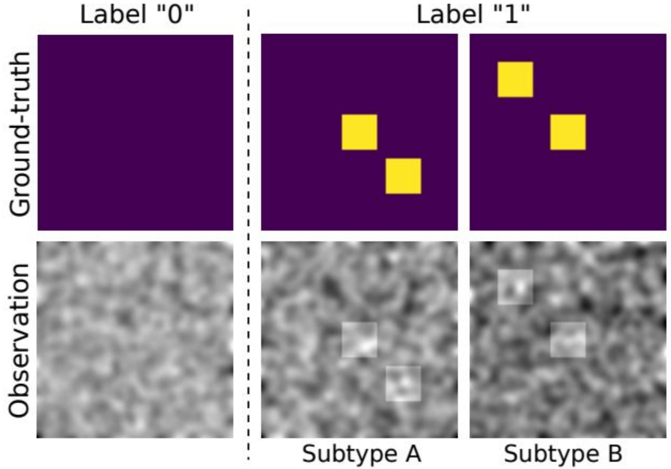


Figure 3. Description of synthetic data. We generated noisy observations from ground-truth effect maps. The dataset contained two categories: A baseline category 0 (e.g. healthy images) and category with an effect (e.g. patient images). The images in category 1 contained one of two subtypes, A or B, which is unknown to the algorithms. A: box in the lower right, B: box in the upper left.

Synthetic Data Experiments

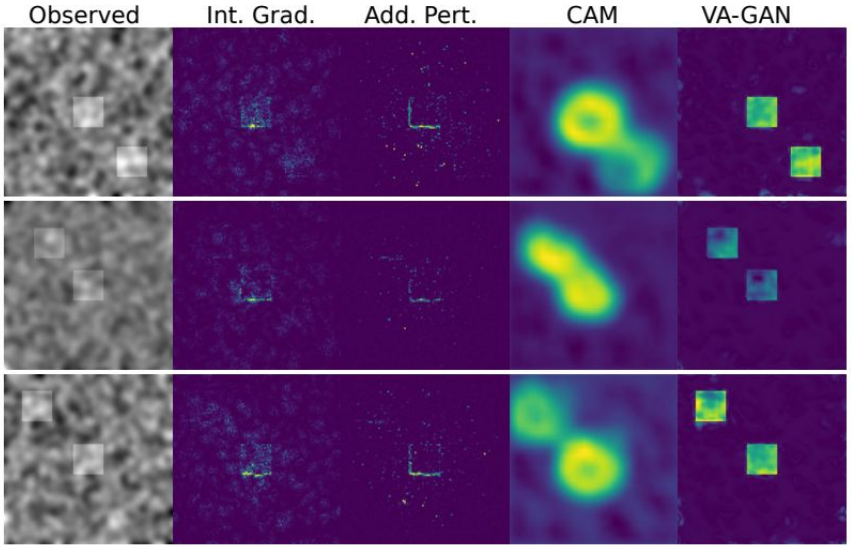


Figure 4. Examples of visual attribution on synthetic data obtained using the compared methods.

Table 1. NCC scores for experiments on synthetic data.

Method	**mean**	**std.**
Guided Backprop [55]	0.14	0.04
Integrated Gradients [56]	0.36	0.11
CAM [67]	0.48	0.04
Additive Perturbation	0.06	0.03
VA-GAN	**0.94**	0.07

Experiments on real neuroimaging data

- Subjects who were diagnosed with MCI during a baseline examination but progressed to AD in one of the follow-up scans.
- We then aligned those images rigidly and subtracted them from each other to obtain an observed disease effect map.
- Training, validation, test: 825, 256, 207 samples

Table 2. NCC scores for experiments on neuroimaging data.

Method	**mean**	**std.**
Guided Backprop [55]	0.05	0.03
CAM [67]	0.09	0.07
Integrated Gradients [56]	0.13	0.05
Additive Perturbation	0.11	0.05
VA-GAN	**0.27**	0.15