DBCE: A Saliency Method for Medical Deep Learning Through Anatomically-Consistent Free-Form Deformations

Joshua Peters†
University of Queensland
joshua.peters@uq.net.au

Léo Lebrat†
CSIRO
QUT

Rodrigo Santa Cruz
CSIRO
QUT

Aaron Nicolson
CSIRO

Gregg Belous
CSIRO

Salamata Konate
CSIRO and QUT

Parnesh Raniga
CSIRO

Vincent Dore
CSIRO

Pierrick Bourgeat
CSIRO

Jurgen Mejan-Fripp
CSIRO

Clinton Fookes
QUT

Olivier Salvado
CSIRO and QUT

Figure 1: Difference between the initial image and the image produced by our generation-based method when applied to a neural network trained to detect Alzheimer’s disease. The blue colour denotes tissue atrophy, and the red colour an increase in tissue density. DBCE outputs the minimal anatomically plausible diffeomorphism flipping one’s network prediction.

Abstract

Deep learning models are powerful tools for addressing challenging medical imaging problems. However, for an ever-growing range of applications, interpreting a model’s prediction remains non-trivial. Understanding decisions made by black-box algorithms is critical, and assessing their fairness and susceptibility to bias is a key step towards healthcare deployment. In this paper, we propose DBCE (Deformation Based Counterfactual Explainability). We optimise a diffeomorphic transformation that deforms a given input image to change the prediction of the model. This provides anatomically meaningful saliency maps indicating tissue atrophy and expansion, which can be easily interpreted by clinicians. In our test case, DBCE replicates the transition of a patient from healthy control (HC) to Alzheimer’s disease (AD). We benchmark DBCE against three commonly used saliency methods. We show that it provides more meaningful saliency maps when applied to one subject and disease-consistent atrophy patterns when used over a larger cohort. In addition, our method fulfils a recent sanity check and is repeatable for different model initialisations in contrast to classical sensitivity-based methods.

1. Introduction

During the last decade, medical Deep Learning (DL) models have surpassed deterministic approaches by delivering faster runtimes and above human-expert performance [42, 60, 35]. Despite these astonishing results, DL-based methods suffer slow adoption rates in healthcare settings, with critics often referring to their lack of tractability or explainability [20, 11]. Indeed, to benefit from a black-box prediction in a critical domain, one would like
to provide additional explanations that have motivated the decision to help the final user in accessing the network decision. Those explanations will first help rationalise and make the decision-making of neural networks (NNs) less opaque. In addition, it could help the end-user identify potential biases or support designing trustworthy and fair algorithms [39, 53].

To this end, numerous methods have been devised to provide the user with an understanding of the “why” through heatmaps that highlight the most critical parts of the input contributing to the prediction. Nevertheless, the quality and “finesse” of these saliency maps are vital for providing additive value to the diagnostic process. It has been demonstrated that for assisting in the grading of diabetic retinopathy, the prediction of a DL model along with a heatmap generated with Integrated Gradient does not provide a more significant benefit over the prediction alone [49]. In addition, recent studies suggest that diffuse saliency maps are prone to user confirmation and automation bias [48, 20], raising further questions on the actual usefulness of saliency maps. In this direction, researchers devised new sanity checks [1, 22, 4], a set of benign tests that we expect saliency methods to pass. Those tests aim to assess the utility and robustness of saliency methods and evaluate their: localisation utility, sensitivity to model weight randomisation, repeatability and reproducibility. Surprisingly, most post-hoc saliency approaches failed to pass those tests [1, 4].

Those recent findings pushed the community to rethink the way of producing saliency maps. Traditionally, sensitivity-based methods look only at what pieces of information already present in the image are used to produce the prediction. Training Data Based Explanation Methods is a new group of regimes that no longer look at an individual image but rather explore the relationships between the image and the training dataset. It provides an explanation based on comparing influential samples, concepts or prototypical parts built from the training dataset [30, 31, 8, 26].

However, relying on the training data has inherent limitations in the case of high-dimensional data and scarce datasets. Recent advances in GPU computing and the progress of generative models allow the creation of compelling counterfeit data. The ability to mimic one’s dataset distribution allowed one to consider counterfactual approaches and explore the generation of explanation by producing new counterfactual examples [57, 19]. Generation-based methods provide a rich insight. It allows exploring the “what if” scenarios through the generation of new plausible examples close to the original image but for which the neural network predicts a different class. Indeed if the classification output is incorrect, this method allows grasping which parts of the image are needed to be changed to correct the prediction. In the case of medical imaging, this approach exhibits to a practitioner which medical images would have yielded a different diagnosis, diminishing the deep neural network’s (DNN) opaqueness. This technique provides a profound insight into the decision boundary of a DNN by providing realistic data of where the decision is being flipped.

In this paper, we propose **DBCE** (Deformation Based Counterfactual Explanation), a generation-based method which given an input image and a black-box Deep Neural Network (DNN), produces a delusive image by optimising a regular deformation on the input image. From DBCE’s deformation, we derive a pixel attribution map (saliency map) derived from the norm of the deformation; we compare this method against classical post-hoc explainability methods. Further, we propose a more instructive visualisation technique based on the difference between the original input and the generated delusive image. We access the repeatability of our technique for a fixed method and architecture but different checkpoints [1, 4]. We then evaluate the intra-class (same diagnosis) repeatability of our method across multiple patients. Finally, we access the ill-posedness of DBCE for deformations supported by a single anatomical zone; we use this result to demonstrate the robustness of lightweight deep Convolutional Neural Networks (CNNs) against localised alterations of the input image. These results highlight the robustness of this architecture [23].

### 2. Related works

Saliency maps for medical imaging have been produced in several ways. They can be derived from the model’s architecture. For example, the attention matrices of a vision Transformer highlight the sub-regions of a medical image that it deems most important [21, 12]; Generative approaches can provide the explanation of a decision’s landscape through displacement in the latent space [57, 10]. However, these methods have several disadvantages: they are computationally intensive, especially for three-dimensional images, and can be challenging to train, particularly on small datasets.

Interpretation of a model’s prediction can be studied through its approximation. This can be achieved locally by discretizing a complex model with a fully explainable one [45]. However, it has been found that the fidelity of this surrogate model can be brittle [3, 58].

Saliency maps can be generated by post-hoc methods, which require a trained model and input samples. There are two main categories of post-hoc methods: sensitivity-based and perturbation methods.

Sensitivity-based methods encompass both gradient-based methods [55, 68, 61, 52, 54, 59, 7, 29] and contribution propagation methods [6, 37]. Notwithstanding their computational affordability, few of these methods require access to intermediate layers [6, 52] or require architectural
modifications of the CNN [68, 61]. However, their interpretation can be difficult and subjective [1, 22, 4].

Perturbation methods aim to find which localized regions contribute most to the prediction by leveraging masking or blurring of the input image [40, 18, 64, 32]. Whilst such perturbations could be appropriate for natural images (where occlusion can naturally occur), they may not be suitable for three-dimensional structural images—as such perturbations would produce unrealistic images. To amend this, three-dimensional perturbation methods often require anatomical segmentation—which is task-dependent and requires labour-intensive annotation [64].

Counterfactual methods [66, 13] are crafted to produce minimal information that will tamper with the initial prediction of a neural network. In the Computer Vision literature, those methods are also known as adversarial perturbations [38, 44] notwithstanding being robust to a physical-world context [33, 16] the insight provided by such methods remains limited. Indeed, those methods results are not sparse and interpretable by humans. Most of the time, examples provided by those methods are artificial and unrealistic to the initial training dataset. Recently, those methods have been successfully revisited using generative approaches with discriminative losses [57, 19]. However, those approaches remain very challenging for three-dimensional, limited data. Our method is situated at the interface of both counterfactual and perturbation based-methods and leverages the benefits from both.

The proposed approach does not require domain knowledge and is model agnostic. It optimises over a set of diffeomorphisms; these transformations do not annihilate the output image’s verisimilitude and allow the generation of high-resolution counterfactual examples. Moreover, the resulting optimisation problem is numerically affordable in a few seconds using a GPU and precludes resorting to computationally expensive generative models whilst generating anatomically plausible images.

This paper evaluates this idea using 3D volumetric structural MR images, with the following assumptions:

- The disease continuum can be modelled by smooth and invertible mappings.
- The mapping from the healthy control to the disease group can be generated using a sufficiently refined Free-Form Deformation (FFD).

The rationale behind generating a new anatomically plausible image using smooth deformations is motivated by numerous medical imaging pipelines that utilise those diffeomorphic mappings to compare, segment and aggregate different measurements between patients [43, 17]. The particular choice towards FFD to parameterise diffeomorphisms is guided by their conciseness, allowing for easier optimisation problems over a set of diffeomorphisms.

3. Method

Given a DNN $f_0$, DBCE smoothly deforms an input image $x$ to produce a counterfactual example $\tilde{x}$. The general flowchart diagram is presented in Figure 2. More specifically, consider a trained DNN $f_0 : \mathbb{R}^{H \times W \times D} \mapsto \mathbb{R}$ and an input image $x \in \mathbb{R}^{H \times W \times D}$ for which one wants to provide a saliency map for the decision $p$. Given the input $(f_0, x)$, DBCE seeks to optimize a smooth deformation $T_\theta(\bullet)$ to produce a counterfactual prediction $f_0(\tilde{x}) = p$ from the deformed image $\tilde{x} = T_\theta(x)$.

A byproduct of using this intuitive set of deformations for medical images is that one can efficiently compute the distance between two anatomies as the energy of the deformation provided by $T_\theta$.

Parameterisation of the deformation. To produce smooth deformations that are invertible and anatomically plausible, we make use of a FFD parameterised by a grid of points $\Phi \in \mathbb{R}^{N \times N \times N \times 3}$ [47]. More generally, $T_\theta(\bullet)$ is the re-sampled digital image determined by the deformation vector field $v_\Phi : \mathbb{R}^3 \mapsto \mathbb{R}^3$ which is defined as,

$$ u_\Phi(x_1, x_2, x_3) = \text{Id}_{\mathbb{R}^3}(x_1, x_2, x_3) + u_\Phi(x_1, x_2, x_3), \quad (1) $$

with Id the identity map, and with $u_\Phi$ defined as,

$$ u_\Phi(x_1, x_2, x_3) = \sum_{l,m,n=0}^{N-1} \beta_l(u)\beta_m(v)\beta_n(w)\phi_{l+m+n}(u,v,w), \quad (2) $$

where $(i, j, k)$ are the local indices within the FFD grid and $(u, v, w)$ their relative position. For instance along the first dimension $i = \lfloor \frac{x_1}{N-1} \rfloor$ and $u = \frac{x_1}{N-1} - (i + 1)$. $\beta_i$ denotes the polynomial decomposition of the third order B-spline function [34] with,

$$ \beta_0(t) = \frac{1}{6}(1-t)^3 $$

$$ \beta_1(t) = \frac{1}{6}(3t^3 - 6t^2 + 4) $$

$$ \beta_2(t) = \frac{1}{6}(-3t^3 + 3u_t^2 + 3t + 1) $$

$$ \beta_3(t) = \frac{1}{6}t^3. $$

DBCE’s optimisation algorithm. Given the aforementioned FFDs, we can now describe the optimisation problem that DBCE seeks to solve,

$$ \Phi^* = \arg \min_{\phi} \sign(f_0(x)) f_0(T_\theta(x)) $$

$$ \text{subject to} \quad f_0(x), f_0(T_\theta(x)) \geq 0 $$

$$ + \sum_{q=1}^{3} \lambda_q R_q(\Phi). \quad (Opt-DBCE) $$

1961
Figure 2: Left: DBCE method: given a Deep Neural Network (DNN), an input image $x$ and its prediction $p$, DBCE seeks to optimize the FFD grid-points $\Phi$ such that when the warped image $\tilde{x}$ is fed to the network the prediction is swapped. Right: Difference between the original image and the image perturbed by DBCE.

Suppose that $f_\theta : \mathbb{R}^{H \times W \times D} \mapsto [-1, 1]$ is a binary classifier occluding from the regularisation terms $R$, the solution is met when $f_\theta(T_\theta(x)) = 0$ for a modified image $T_\theta(x)$ on the decision boundary of $f_\theta$. We compute (Opt-DBCE) using a gradient descent approach, as described by Algorithm 1.

**Algorithm 1 DBCE Algorithm**

**Input** Image $x$, model $f_\theta$

**Output** Image $\tilde{x}$

Initialise $i \leftarrow 0$, $x_0 \leftarrow x$, $\Phi_0 \leftarrow 0_{\mathbb{R}^{N \times N \times N \times 3}}$

\* $\bullet_i$ denotes the variable $\bullet$ at the $i$-th iteration.

While $f_\theta(x_i)f_\theta(x_i) > 0$

\begin{align*}
\mathcal{L}_i & \leftarrow \text{sign}(f_\theta(x))f_\theta(x_i) + \sum_{q=1}^{3} \lambda_q R_q(\Phi_i) \\
s_i & \leftarrow \nabla_{\Phi_i} \mathcal{L}_i \\
\Phi_{i+1} & \leftarrow \Phi_i - \tau s_i \\
x_{i+1} & \leftarrow T_{\Phi_{i+1}}(x_i) \quad \triangleright \text{Gradient step} \\
i & \leftarrow i + 1
\end{align*}

**end while**

**return** $\tilde{x} = x_i$

**Penalties.** To meet the assumptions presented at the end of Section 2, one has to enforce constraints on the deformation $T_\theta$. This is implemented by adding penalty terms to the loss function.

The first restriction considered is the local invertibility of the mapping provided by the FFD. We borrow the quadratic penalisation proposed by [9],

$$R_1(\Phi) = \frac{1}{N^3} \sum_{d=1}^{3} \sum_{i,j,k=1}^{N} \left( p(\phi_{i+1,j,k,d} - \phi_{i,j,k,d}; \chi_1^{d_x}, \chi_1^{d_x}) + p(\phi_{i,j,k+1,d} - \phi_{i,j,k,d}; \chi_2^{d_x}, \chi_2^{d_x}) + p(\phi_{i,j,k+1,d} - \phi_{i,j,k+d}; \chi_2^{d_x}, \chi_2^{d_x}) \right)$$

with $p$ a quadratic penalty function defined as,

$$p(t; \chi_1, \chi_2) = \begin{cases} 
\frac{1}{2}(t - \chi_1)^2 & \text{if } t < \chi_1 \\
\frac{1}{2}(t - \chi_2)^2 & \text{if } t > \chi_2 \\
0 & \text{otherwise.}
\end{cases}$$

This ensures that the deformation $v_\Phi$ is invertible and averts the creation of foldings and singularity points. It is here where anatomical information is lost and artifacts can be created.

The second constraint is an elementary $L^1$-regularisation. For height $H$, width $W$ and depth $D$ of the image, we enforce the sparsity of the transformation via,

$$R_2(\Phi) = \frac{1}{HW D}\|u_\Phi\|_1.$$  \hspace{1cm} (3)

Finally, we enforce the support of $u$ to be restrained to $\Omega \subset \mathbb{R}^3$,

$$R_3(\Phi) = \frac{1}{N^3}\|(1 - M_\Omega) \odot \Phi\|_1,$$  \hspace{1cm} (4)

where $\odot$ denotes the Hadamard product and $M_\Omega$ is a discrete binary mask down-sampled to the grid points’ resolution. This term allows us to localise the transformation.
Figure 3: Saliency maps for each method of an axial and a sagittal view for a single prediction. Percentage of total saliency map averaged for all of the scans present in the experimental split and correctly classified as HC.
to specific anatomic regions of the input image. Throughout Section 4 and 5 (except for Section 5.2) we restrict Ω to the patient’s brain. This ensures that deformations occur in parts of the patient’s anatomy that are susceptible in changing the diagnosis.

**Implementation details.** DBCE is model agnostic and is not restricted to only convolutional architectures [68, 52, 32] but is available for any differentiable architecture. We efficiently implement the FFD by leveraging strided transposed convolutions resulting in a small memory footprint of 4.92GiB and a fast runtime of 0.256 ± 0.008s per gradient step. Algorithm 1 converges in 32.9 ± 13.4 iterations (∼8.45s) for a 256 × 256 × 256 image.

4. Experiments

Early diagnosis of Alzheimer’s Disease (AD) is a crucial task in neuroscience. The sooner the disease is detected, the more time is given for medical intervention and improving the patient’s quality of life [2]. Promising deep-learning-based approaches have been devised for the early detection of AD [27, 63]. However, the lack of interpretation for their decisions slows down their implementation in clinical practice, and it is difficult to derive new knowledge from those models. In this section, we evaluate DBCE for a DNN, which, given a three-dimensional T1w structural image, predicts if the patient is healthy or has Alzheimer’s disease. The architecture is inspired by a lightweight CNN [23]. We train this model using early stopping on an augmented version of the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset with a train/validation/test split of 776/342/520 images. We follow the procedure described in [67] to prevent any data leakage. More details on the architecture and dataset splits are provided in the supplementary materials.

4.1. A new visualisation technique provided by DBCE

The counterfactual images ˜x provided by Algorithm 1 are very close to the initial images x. Indeed, the global deformation of the FFD defined in Equation (1) is set to be the identity. Moreover, the regularisation term in Equation (3) ensures the deformation is both sparse and has small energy. As a result, the difference of x − ˜x for a given patient can be used to display voxel intensity variations between the original and the delusive image accurately. In addition, the brain structures of adversarial examples are aligned to the original patient, which accord with the use of Voxel-Based Morphometry approaches [5]. This method is also suitable for “gifplanation” [10]. We attached such videos in our supplementary material. As depicted in Figure 6, DBCE appears to modify the size of the ventricles and the thickness of the temporal cortices and the hippocampal formation.

4.2. Qualitative analysis of DBCE’s saliency map and comparison to pre-existing methods

In this section, we evaluate the saliency maps derived from the local norm of the transformation produced by DBCE against the Gradient method [56], DeconvNet [68], and GradCam [51]—three standard techniques for post-hoc interpretability.¹

In Figure 3, we display saliency maps for each of the methods when considering images that are correctly classified as a Healthy Control (HC). Saliency maps for DeconvNet and Gradient are very scattered and diffused, thus challenging to interpret. On the contrary, maps issued from GradCam are very evasive, highlighting a large zone that leaves room for interpretation bias. Finally, we propose to compare the methods by computing the averaged percentage of energy present in the patient’s cortical sub-structures using FreeSurfer’s atlas [14]. The saliency map derived from DBCE indicates that most of the deformations are taking place in the temporal lobes, which are discriminating regions for diagnosing AD [28].

5. Accessing the trustworthiness of the proposed method

The validation of saliency maps is a non-trivial topic as no ground truth exists, and the results could depend on the neural network architecture used. In this section, we will present different experiments that assess the reliability and usefulness of the saliency maps produced by DBCE.

5.1. Saliency maps repeatability

One desirable characteristic of a saliency method is its repeatability. When training the same architecture with different initialisations, one hopes to converge towards models that have learned the same patterns in the data [4]. Different saliency maps produced for different models should be comparable for the same input image. To access this behaviour, we propose to compare the resulting saliency maps for four different initialisations (repetitions) of the investigated network [23], and for saliency maps produced by different saliency methods [56, 68, 51, 59, 7, 29].

¹Implementation recovered from TorchRay https://facebookresearch.github.io/TorchRay/
Following [1], we use the structural similarity index (SSIM) and the Pearson’s correlation of the histogram of gradients (HOGs) to compute the repeatability of two saliency heatmaps for two different models on the same image. We used the implementations provided by scikit-image toolbox [65] for the computation of the HOGs. We computed the gradient deep-wise with $(16, 16)$ pixels per cell and concatenated all of the resulting histograms before computing the Pearson’s correlation. The quantitative comparison of the method is reported in Figures 4 and 5.

According to [4], intra-architecture repeatability is achieved for DBCE with a structural similarity index greater than $0.5$ ($0.745 \pm 0.039$).

5.2. Robustness to localised deformations

We propose to evaluate the robustness of a prediction to deformations carried out on individual anatomical substructures. Using the segmentation masks of FreeSurfer [14], we modify Algorithm 1 to restrict the deformation sequentially to every parcellation from this atlas. In Figure 6, we report the success rate in producing such a deformation that changes the DNN’s prediction from Healthy Control (HC) to Alzheimer’s Disease (AD).

Surprisingly, few of these deformations manage to change the prediction initially made by the neural network. This experiment suggests that the decision taken by a lightweight CNN [23] stems from a combination of different image features that are spatially disjoint. Similarly, this result indicates that smoothly deforming local features in isolation (a scenario that is not anatomically plausible and not described in the dataset) is not likely to tamper with the prediction of a CNN. This result could seem confusing in comparison to ultra-localized adversarial attacks where one-pixel change affects the whole prediction [62]. Nevertheless, one has to recall that by construction, FFD will continuously change the value of the input image, whereas flipping a well-chosen pixel value can create significant discontinuity that can brutally affect the value of the downstream feature maps. In addition, DBCE highlights an asymmetry in the decision made by the DNN, which might reflect an asymmetry in atrophy due to AD or a bias in the model [46, 64].

5.3. Intra-class reproducibility

The display of individual examples does not allow for a general grasp of one’s network behaviour. A pleasant characteristic of numerous medical conditions is that for a given diagnosis, the causes of the disease should be similar across the whole cohort. For the particular case of AD, the disease’s development is associated with atrophies of the temporal cortices and the hippocampal areas [15, 41]. In order to evaluate across all of the images with a similar prognosis, we propose to visualise the averaged absolute difference between the initial image and the deformed image produced...
by DBCE. To unify this result across a broad range of patients, we perform a non-rigid registration [36] towards an averaged brain-atlas [24].

As depicted in Figure 7, the absolute difference averaged across 235 patients tends to be localised in the temporal and the ventricle regions. The frontal, parietal and occipital areas are not highlighted by DBCE, which appear to be coherent with the current understanding of AD [25, 50, 28].

In addition, this experiment illustrates the robustness of the presented method. Indeed, for different anatomies and MR images, the presented method tends to consistently highlight disease-specific anatomical areas. In our supplementary material, we provide additional experiments on the robustness of our approach to additive Gaussian noises.

6. Conclusion

The adoption and advancement of deep-learning methods applied to healthcare applications will hinge on researchers’ efforts to provide robust analysis and explanation methods that reduce the opaqueness of DNNs. This paper introduces DBCE, a generation-based interpretability method based on smooth deformations of the input image, available in 3D, and which does not require additional training data or the use of generative models. In contrast, the generation of anatomically plausible images relies on the resolution of an optimisation problem. To ensure deformations are anatomically plausible, we derive three penalty functions that allow one to tune the sparsity, localisation and invertibility of those deformations. Adjacently, we introduce two visualisation techniques for this method that monitors the voxel changes in intensity between the original and counterfactual image, or the local energy of the deformation. Finally, qualitative and quantitative tests are described to evaluate the usefulness and trustworthiness of our approach.

7. Compliance with Ethical Standards

This research was approved by CSIRO ethics 2021 068 LR

8. Acknowledgements

This work was funded in part through an Australian Department of Industry, Energy and Resources CRC-P project between CSIRO, Maxwell Plus and I-Med Radiology Network.

References


