Dual Decomposition of Convex Optimization Layers for Consistent Attention in Medical Images

Tom Ron [*1] **Michal Weiler-Sagie** [2 3] **Tamir Hazan** [*1]

Abstract

A key concern in integrating machine learning models in medicine is the ability to interpret their reasoning. Popular explainability methods have demonstrated satisfactory results in natural image recognition, yet in medical image analysis, many of these approaches provide partial and noisy explanations. Recently, attention mechanisms have shown compelling results both in their predictive performance and in their interpretable qualities. A fundamental trait of attention is that it leverages salient parts of the input which contribute to the model's prediction. To this end, our work focuses on the explanatory value of attention weight distributions. We propose a multi-layer attention mechanism that enforces consistent interpretations between attended convolutional layers using convex optimization. We apply duality to decompose the consistency constraints between the layers by reparameterizing their attention probability distributions. We further suggest learning the dual witness by optimizing with respect to our objective; thus, our implementation uses standard backpropagation, hence it is highly efficient. While preserving predictive performance, our proposed method leverages weakly annotated medical imaging data and provides complete and faithful explanations to the model's prediction.

1. Introduction

Since the emergence of artificial intelligence, there has been great interest in machine learning algorithms' reasoning processes. In high-stakes domains, such as medical image analysis, understanding the inner workings behind a model's decision is crucial for building trust and accountability between end-users and algorithms. The field of explainable artificial intelligence (XAI) generally addresses two notions regarding model interpretation: the *"how"* and the *"why"*. The former is often focused on the model's architecture and optimization process, which grew to be challenging as models became complex. By contrast, the latter can be viewed from a data-centered perspective: which particular attributes of the input inclined the model towards a specific decision.

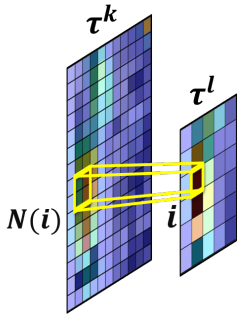

Figure 1. We propose a multi-layer visual attention mechanism for obtaining comprehensive explanations in medical image recognition. One layer is closer to the input and it obtains a detailed attention of the input, while the other is closer to the output and it obtains a detailed attention of the prediction. The decision is performed only on attended areas in the image and consequently we encourage comprehensive explanation. Using a spatial marginalization layer, we enforce consistency between the two attention units to ensure a consistent explanation.

To this end, a magnitude of explainability methods has been proposed in recent years, including local linear approximations, gradient-based methods, post-hoc explanations and attention-based explanations: A theoretically solid class of explainers are based on local linear approximations (Ribeiro et al., 2016; Sundararajan et al., 2017; Bach et al., 2015; Shrikumar et al., 2017). These methods often require a reference point for computing an explanation for a specific input. However, finding an appropriate reference point for each input is unwieldy, particularly in neural networks with non-linear activations (Samek et al., 2021). Gradient-based meth-

[*]Equal contribution [1]The Faculty of Industrial Engineering and Management, Technion Institute of Technology, Haifa, Israel [2]Department of Nuclear Medicine, Rambam Health Care Campus, Haifa, Israel [3]Rappaport Faculty of Medicine, Technion Institute of Technology, Haifa, Israel. Correspondence to: Tom Ron <tomron27@campus.technion.ac.il>, Tamir Hazan <tamir.hazan@technion.ac.il>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

ods (Sundararajan et al., 2017; Srinivas & Fleuret, 2019; Smilkov et al., 2017; Shrikumar et al., 2017; Selvaraju et al., 2017) are another popular class of explainers, but they often suffer from volatility (Balduzzi et al., 2017) or output coarse explanations. In addition, recent analysis has shown that in computer vision, gradient-based explanations degrade into edge-detectors (Adebayo et al., 2018) and exhibit other pathologies (Feng et al., 2018). The majority of methods are typically applied post-hoc (Bach et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017; Smilkov et al., 2017; Srinivas & Fleuret, 2019; Selvaraju et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017). That is, given a trained model and an input example, they observe the propagation of input throughout the model and process this observation into an explanation.

Unfortunately, such explanation methods have inherent weaknesses upon application in medical image analysis. As opposed to most tasks in natural image recognition, detecting pathologies in medical images typically requires expertise and years of training for humans. Features such as color, texture, and shape in the medical image landscape are often intricate and subtle to discern. Volatility and incompleteness of post-hoc methods can easily exacerbate and produce noisier explanations compared to the natural image setting. In addition, as modern-day convolutional nets are optimized solely for predictive performance, analysis of post-hoc explanations typically reveals that the models focus on the top-most discriminative features of the input. While this reliance on a small subset of features might provide plausible explanations for natural images, medical image analysis imposes stricter requirements. Often the medical physician would demand a broader glimpse into the algorithm's reasoning process, i.e., to be presented with *all* the discriminative features available in the input. Our experiments demonstrates that this information is not well tractable using post-hoc explanation methods. In pursuit of this notion, Rudin (2019) argued against using post-hoc explanations for black-box models (particularly in high-stakes settings) and advised using interpretable models instead. Our work is motivated by such criteria, where we use intrinsic attention to produce reliable and complete interpretations.

A prominent class of machine learning algorithms is attention-based models (Bahdanau et al., 2014; Vaswani et al., 2017; Devlin et al., 2018). Alongside high predictive performance, these models admit transparency through their attention weight distributions. Recent works have discussed whether attention distributions can be considered as explanations to model prediction (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Jain et al., 2020; Yu et al., 2019), resolved with the notions of plausible, faithful and comprehensive explanations. These works are focused on natural language processing while our work aims at learning comprehensive explanations using attention for visual models, i.e, explanations that contain all the relevant information for the model's prediction.

In this work, we propose a multi-layer visual attention mechanism for obtaining comprehensive explanations in medical image recognition. To enforce comprehensive explanations we learn two attention layers: one layer is closer to the input and it obtains a detailed attention of the input, while the other is closer to the output and it obtains a detailed attention of the prediction. The decision is performed only on attended areas in the image and consequently we encourage comprehensive explanation. Using a spatial marginalization layer, we enforce consistency between the two attention units to ensure a consistent explanation. Using dual decomposition, we demonstrate that our constraint admits an optimal solution that is equivalent to a solution obtained by a convex optimization program, yet it is achieved efficiently by end-to-end backpropagation. We assert the quality of explanations produced by our approach in two experiments, one conducted on a self-curated, novel dataset consisting of 2637 labeled images of ^{18}F-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography (FDG-PET/CT) patient scans.

2. Related work

Explanation by attribution propagation. Attribution propagation methods follow an iterative scheme where a model's output (representing an initial relevance score) is backpropagated and decomposed across layers of the model. Formalized by Montavon et al. (2017), these methods apply a Taylor expansion at each layer of the network, and attribution is repeatedly distributed until reaching the input. Layer-Wise Propagation (LRP) (Bach et al., 2015) backpropagates a prediction score while weighting contributions from each layer in networks using ReLU activations. DeepLIFT (Shrikumar et al., 2017) decomposes the output by computing differences of contribution scores between the activation of each neuron to some reference point. Attribution propagation methods are typically fast to compute and provide interpretation at the pixel level.

Other attribution techniques include saliency and occlusion (Dabkowski & Gal, 2017; Mahendran & Vedaldi, 2016; Simonyan et al., 2014; Zeiler & Fergus, 2014; Zhou et al., 2016), architectural modifications (Zhang et al., 2018) and input perturbation (Fong et al., 2019; Fong & Vedaldi, 2017). In particular, popular methods such as LIME (Ribeiro et al., 2016), Shapley-value sampling (Lundberg & Lee, 2017) and L2X (Chen et al., 2018) offer solid theoretical frameworks for model interpretation, but are computationally heavy as they require multiple evaluations or substantial explanation layers.

Explanation by gradient analysis. Gradient-based methods leverage gradients computed during back-propagation to obtain a relevancy score for the input of each layer or neuron. The magnitude and sign of gradients can be visualized, thus providing an explanation for which parts of a layer's input were most influential to its output. Gradient×Input (Shrikumar et al., 2017) multiplies the input to each layer by the partial derivatives of its output. Ancona et al. (2017) demonstrated that under ReLU activations, a zero-baseline and no biases, LRP (Bach et al., 2015) and DeepLIFT (Shrikumar et al., 2017) are equivalent to Gradient×Input. Integrated Gradients (Sundararajan et al., 2017) assign importance to input features by approximating the integral of gradients with respect to the inputs along a path from a given baseline. SmoothGrad (Smilkov et al., 2017) and FullGrad (Srinivas & Fleuret, 2019) attempt to stabilize gradient-based explanations by averaging, adding Gaussian noise, and other aggregative manipulations. Notably, Class Activation Mapping (CAM) approaches such as GradCAM (Selvaraju et al., 2017) combine the activation and gradients of a given layer, conditioned on a target class. GradCAM provides compelling visual results but usually only applicable to very deep layers, resulting in coarse heatmaps.

Attention and explainability. Since introduced by Bahdanau et al. (2014), attention mechanisms have grown ubiquitous in image recognition (Jetley et al., 2018; Ramachandran et al., 2019; Bello et al., 2019; Hu et al., 2019; Zhao et al., 2018; Hu et al., 2018; Wang et al., 2017) and in multimodal tasks such as visual question answering (Anderson et al., 2018; Yu et al., 2017; Schwartz et al., 2017) and visual dialog (Gan et al., 2019; Schwartz et al., 2019b;a). Oktay et al. (2018) demonstrated the predictive benefits of visual attention in medical image segmentation. In natural language processing, transformers (Vaswani et al., 2017; Devlin et al., 2018) rapidly became the backbone to many state-of-the-art models in machine translation (Wang et al., 2019; Mehta et al., 2020), question answering (Yamada et al., 2020) and sentiment analysis (Jiang et al., 2019; Raffel et al., 2019; Lan et al., 2019). Promising attempts to adjust transformers toward tasks in computer vision have taken place. Dosovitskiy et al. (2020) introduced the Visual Transformer (ViT), which treats patches of the input image as tokens. Although being resource-intensive, it demonstrated competitive results with state-of-the-art convolutional networks in classifying natural images.

Inspired by bottom-up/top-down functions of cognitive processes (Connor et al., 2004), attention mechanisms are often designed to focus on input features that are significant to the model's prediction. Considering whether attention weight distribution can rightfully explain the model's prediction was at the center of a recent discussion. Jain & Wallace (2019) empirically demonstrated that drastically distorting attention weights has little effect on model performance, hence considering these to have explanatory value is spurious. In return, Wiegreffe & Pinter (2019) refined the discussion and considered two types of explanations: plausible and faithful. Plausible explanations are described as such that would seem intuitive to a human being, e.g., a heatmap highlighting *some* discriminative features of a target class. Faithful explanations guarantee that *only* the highlighted features were used in the model's prediction; therefore, they can be causally associated with the prediction. Yu et al. (2019) has further extended these criteria into the notion of comprehensiveness: explanations must contain *all* the relevant information for the model's prediction. Along with the demand of Rudin (2019) to devise models with internal explanations, we follow the faithfulness and comprehensiveness paradigms in the design of our method.

Convex optimization. Convex optimization methods have been recently integrated into deep neural networks. Amos & Kolter (2017) introduced OptNet, a neural architecture that can encode quadratic programs as trainable layers. Agrawal et al. (2019) proposed a new grammar and translation form for convex problems, allowing to automatically differentiate through convex layers using standard back-propagation. Domke (2010), Paulus et al. (2020) and Berthet et al. (2020) offered to compute gradients to non-differentiable operations by perturbation, which in turn can be used to back-propagate with respect to a solution of a convex problem. Nevertheless, differentiation by perturbation and convex optimization layers poses significant computational burdens. The former often involves multiple sampling schemes and generates biased gradients, and the latter is typically solved synchronously while heavily relying on sequential computations. Our work differs from the aforementioned works, as we use convex duality to decompose the optimization and rely on the dual reparameterization to ensure optimality. Moreover, we suggest learning the dual witness within the deep net; thus, we can enjoy the benefits of parallelized GPU operations. Additional overview on machine learning and dual decomposition can be found in (Rush & Collins, 2012; Sra et al., 2012).

3. Explaining decisions on medical images using consistent attention

We learn consistent attention probability distributions to explain machine learning decisions in medical image recognition. Our primary task is learning to classify levels of altered biodistribution of FDG in PET/CT scans while providing comprehensive interpretations to the algorithm's prediction. To encourage comprehensive explanations we restrict the model to rely only on attended neurons while preserving the spatial structure of the image. We use the attention probability model to weigh the amount of visual information that flows through the convolutional layers.

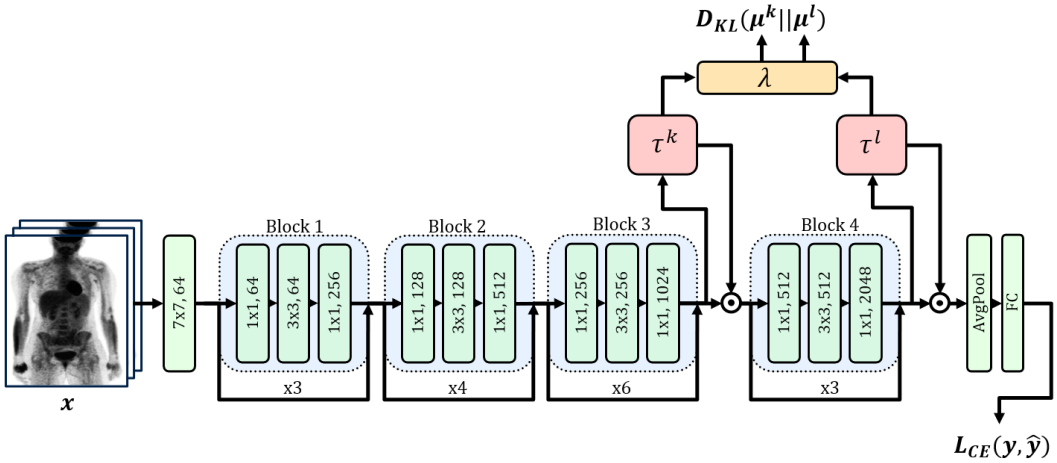

Figure 2. ResNet-50 architecture with attention gates (τ^k, τ^l) and a marginal consistency layer λ. Enforcing the marginalization consistency using convex optimization layer (see Eq. 4) couples the whole block into a single optimization layer. The duality theorem allows to decompose these layers and efficiently enforce this constraint using reparameterization (see Theorem B.1).

A convolutional layer is a 3-dimensional objects of dimension $n \times n \times d$. The $n \times n$ cells corresponds to $n \times n$ sub-windows in the image. The spatial location of each sub-window corresponds to a convolutional neuron and is indexed by $i = 1, ..., n^2$. The i^{th} convolutional neuron is described by the vector $v_i \in \mathbb{R}^d$ which represents the corresponding sub-window in the image. In the following, we refer to the k^{th} convolutional layer by an additional index, i.e., the dimension of the k^{th} layer is $n_k \times n_k \times d_k$ and the i^{th} sub-window representation is denoted by $v_i^k \in \mathbb{R}^{d_k}$.

We learn an attention probability distribution over intermediate convolutional layers to weigh the information of their i^{th} neuron. Consequently, we are able to emphasize the information in the relevant sub-windows using the learned attention probability distribution. Formally, for the k^{th} convolutional layer, we learn a probability distribution τ_i^k over the $n_k \times n_k$ sub-windows of the input image. Given $v_i^k \in \mathbb{R}^{d_k}$, we learn τ_i^k by first learning its potential function

$$\phi_i^k = \left\langle u_i^k, \text{relu}\left(\frac{U_k v_i^k}{||U_k v_i^k||}\right)\right\rangle \tag{1}$$

where $U_k \in \mathbb{R}^{d_k \times d_k}$ and $u_i^k \in \mathbb{R}^{d_k}$ are learnable parameters.

The attention probability distribution is set to be its potential function softmax, namely,

$$\tau_i^k = \frac{e^{\phi_i^k}}{\sum_{j=1}^{n_k^2} e^{\phi_j^k}}. \tag{2}$$

Intuitively, this attention probability distribution assigns high probabilities to sub-windows that are important in the learner's decision and assigns low probabilities to sub-windows that do not contribute to the decision. By doing so, visualizing the attention probability distribution introduces spatial information to the medical physician, accounting for areas in the image that the learner relies on in its decision making (see Fig. 1).

Our spatially attended embedding, $\hat{v}_i^k$, of the k^{th} layer combines both the semantic information in the image as well as the spatial information about the location of the attention distribution. The semantic information is represented by the vectors $v_i^k \in \mathbb{R}^{d_k}$ while the spatial information is represented by the attention probability distribution τ_i^k. We combine these two components using the pointwise multiplication: $\hat{v}^k = \tau^k \odot v^k$, which is defined by

$$\hat{v}_i^k = \tau_i^k \cdot v_i^k. \tag{3}$$

The dimension of the attended embedding $\hat{v}_i^k$ is identical to the dimension of the image embedding v_i^k. Intuitively, the attention probability distribution τ_i^k has high values for relevant i^{th} sub-windows of the image, i.e., sub-windows that contain semantic information for the medical diagnosis task. Equivalently $\tau_i^k \approx 0$ for irrelevant sub-windows. Hence, the attended embedding in Eq. (3) consists of the vector $\hat{v}_i^k \approx 0$ whenever $\tau_i^k \approx 0$, i.e., for sub-windows that are irrelevant for the model's decision in the k^{th} layer. Equivalently, for semantically meaningful sub-windows, the attended embedding is similar to the original image embedding, i.e., $\hat{v}_i^k \approx v_i^k$. This embedding highlights the spatial locations of the semantically meaningful sub-windows and populates them with the respective semantic information of the image. In addition, the embedding restricts the decision of the deep net to rely on sub-windows for which their corresponding attention probability distribution is high. The attention probability distribution, in turn, provides meaningful spatial information to the medical physician interpreting

the scan and points towards the area that the decision relied on. Thus, it serves as a faithful visual explanation for the algorithm's decision in the image.

3.1. Consistent Attention

To enforce comprehensive explanations we learn two attention layers in a standard convolutional nets (e.g., ResNet-50 (He et al., 2016)). Due to the convolutional net structure, these attention probability distributions play different roles in explaining network prediction. The first layer, τ^k, is closer to the input therefore it better represents variability in the data, e.g., gradients in the image. Also, a convolutional layer that is closer to the input has more convolutional neurons and therefore has a higher resolution with respect to the input image. The second layer, τ^l for $k < l$, is closer to the output and provides semantic information about the predicted label.

We suggest to use convex optimization to obtain consistent attention across τ^k, τ^l. We use the spatial relations between layers to enforce marginalization consistency constraints between τ^k and τ^l, as illustrated in Fig. (2). We define $N(i)$ to be the set of indexes $j \in N(i)$ for which τ^k_j corresponds to τ^l_i. In Fig. (1) we can see that $j \in N(i)$ whenever j and i refer to the same sub-window in the image. To guarantee consistent attention, we would like to enforce the marginalization constraint $\sum_{j\in N(i)} \tau^k_j = \tau^l_i$ for every i. To do so, we infer consistent probability distributions μ^k, μ^l that are closest to τ^k, τ^l respectively and satisfy the marginalization constraints. We use the KL-divergence $D_{KL}(\mu||\tau) = \sum_i \mu_i \log(\mu_i/\tau_i)$ to measure the similarity between μ and τ. Thus, the consistent attention probability model is the output of the following convex program:

$$\begin{aligned} &\arg\min_{\mu^k,\mu^l} D_{KL}(\mu^k||\tau^k) + D_{KL}(\mu^l||\tau^l) \\ &\text{s.t. } \mu^k, \mu^l \text{ are distributions, and } \mu^l_i = \sum_{j\in N(i)} \mu^k_j \end{aligned} \tag{4}$$

Unfortunately, it is computationally challenging to solve this program as a convex optimization layer as it couples all parameters from layer k to layer l by the same convex program (see Fig. 2).

3.2. Dual decomposition and primal-dual optimality conditions

We propose to use strong duality to decompose the optimization that couples the k^{th} layer and the l^{th} layer. Lagrange duality theorem relies on Lagrange multipliers λ that reparameterize the primal variables τ^k and τ^l in order to obtain the optimal solutions μ^k, μ^l of the primal program in Eq. (4). Our reparameterization is attained by relating a Lagrange multiplier λ_i to each marginalization constraint $\mu^l_i = \sum_{j\in N(i)} \mu^k_j$. We use the enhanced Fritz-John condition (Mangasarian & Fromovitz, 1967) to enforce the probability distribution constraints for μ^k, μ^l implicitly. Therefore, we are able to attain the following reparameterization for the optimal solution of Eq. (4):

Theorem 3.1. *The probability distributions μ^k, μ^l are the optimal solutions to the primal program in Eq. (4) if (i) $\mu^l_i = \frac{\tau^l_i e^{\lambda_i}}{\sum_s \tau^l_s e^{\lambda_s}}$; (ii) $\mu^k_j = \frac{\tau^k_j e^{-\lambda_i}}{\sum_t \sum_{s\in N(t)} \tau^k_s e^{-\lambda_t}}$; and (iii) $\mu^l_i = \sum_{j\in N(i)} \mu^k_j$, namely, the parameterized versions μ^l, μ^k agree on their marginal distributions.*

Proof. See Section 3.2 of the appendix. □

Strong duality allows us to decompose the constraint across the k^{th} and the l^{th} layers using reparameterization. That is, whenever we are able to find a λ for which the two distributions $\frac{\tau^l_i e^{\lambda_i}}{\sum_s \tau^l_s e^{\lambda_s}}$ and $\frac{\tau^k_j e^{-\lambda_i}}{\sum_t \sum_{s\in N(t)} \tau^k_s e^{-\lambda_t}}$ agree on their marginals, then these distributions are guaranteed to be primal optimal and to propagate along the deep net. Summarizing these conditions, we search for the following reparameterized condition

$$\sum_{j\in N(i)} \frac{\tau^k_j e^{-\lambda_i}}{\sum_t \sum_{s\in N(t)} \tau^k_s e^{-\lambda_t}} = \frac{\tau^l_i e^{\lambda_i}}{\sum_s \tau^l_s e^{\lambda_s}}. \tag{5}$$

The above reparameterization use the dual variables λ to guarantee the primal-dual optimality conditions in Theorem B.1. Hence λ serves as a dual witness to the optimality of the primal variables μ^l, μ^k in Eq. (4) if their reparameterization satisfy their marginalization constraints.

3.3. Learning the dual witness

Unfortunately, the reparameterized constraint in Eq. (5) needs to hold for any data point in training and in particular in any batch operation. While this can be done by projected stochastic gradient descent, this is computationally unappealing. Instead, we suggest learning λ for every pair of τ^k, τ^l. We do so by introducing a shallow layer, parameterized by λ, which encodes the two counterparts of Eq. (5). We then enforce the constraint by adding a penalty function to our optimization objective

$$\mathcal{L}_{CE}(y,\hat{y}) + \alpha D\Big(\frac{\tau^l_i e^{\lambda_i}}{\sum_s \tau^l_s e^{\lambda_s}}, \sum_{j\in N(i)} \frac{\tau^k_j e^{-\lambda_i}}{\sum_t \sum_{s\in N(t)} \tau^k_s e^{-\lambda_t}}\Big) \tag{6}$$

Where $\mathcal{L}_{CE}(\cdot,\cdot)$ is the categorical cross-entropy loss and $D(\cdot,\cdot)$ is a divergence function, e.g., KL-divergence, which encourages the reparameterization condition in Eq. (5). The positive number $\alpha > 0$ controls the loss incurred when the reparameterization condition is not satisfied.

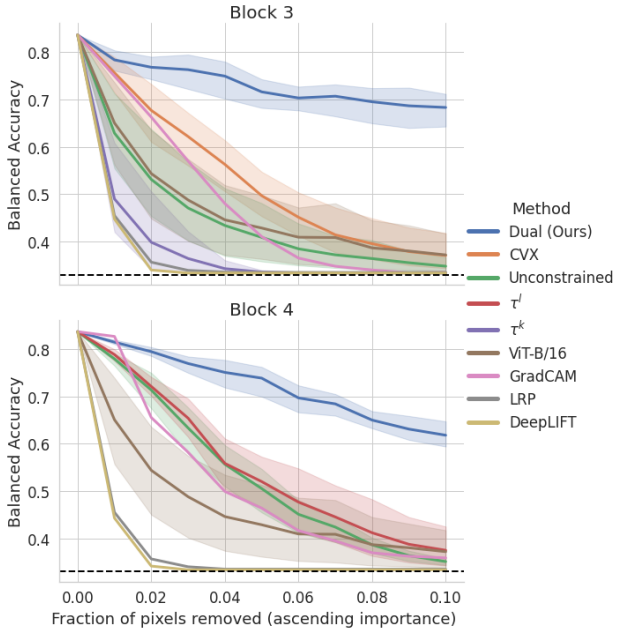

Figure 3. Input perturbation curves on FDG-PET-MUAB dataset. Conceivably, faithful explanations only highlight relevant regions in the input image; thus, removing irrelevant regions should have a lesser impact on downstream accuracy. Compared to the baselines, the classifier demonstrates a significantly lesser decay in accuracy when removing irrelevant pixels according to our *Dual* approach.

4. Experiments

We evaluate the explanatory value of the attention distribution obtained by our method in two experiments on two medical imaging datasets. In each experiment, we train a vanilla CNN model and extract attribution maps using various baselines. In addition, we compare our visual explanations to attention weights obtained from a fine-tuned visual transformer (ViT-B/16) which was pretrained over ImageNet (Wightman, 2019). We divide the baselines into three subgroups: gradient-based (GradCAM (Selvaraju et al., 2017); Gradient×Inputs (Shrikumar et al., 2017)), attribution propagation-based (DeepLIFT (Shrikumar et al., 2017); LRP (Bach et al., 2015)) and attention-based (ViT-B/16 (Dosovitskiy et al., 2020); single-gate and unconstrained attention (Eq. 3); CVX attention (Eq. 4); dual decomposition of consistent attention (Eq. 6)). For methods that are class dependant (Shrikumar et al., 2017; Selvaraju et al., 2017), we condition the attribution on the ground truth label. For methods that provide channel-level attribution (Shrikumar et al., 2017; Bach et al., 2015), we perform 0-1 normalization on each channel and sum attributions across channels. We compare attention attribution maps from several convolutional layers to the class activation maps of the equivalent layers in (Selvaraju et al., 2017), to sum-pooled attribution maps in (Shrikumar et al., 2017; Bach et al., 2015) and to an average of attention heads weights in (Dosovitskiy et al., 2020). Our models and baselines were implemented using PyTorch (Paszke et al., 2019) and Captum (Kokhlikyan et al., 2020) and were trained on a single Nvidia GeForce RTX 2080TI GPU.

Table 1. Area under the curve (AUC) for input perturbation results, FDG-PET-MUAB dataset. We compute the area under the curves depicted in Fig. (3) for a full range of fractions (0.0-1.0). Our *Dual* approach drops to random accuracy when roughly 60% of the input pixels are removed (by order of importance).

Method	Block 3	Block 4
DeepLIFT	20.37	20.37
LRP	20.48	20.48
GradCAM	21.52	21.90
ViT-B/16	21.63±1.17	21.63±1.17
τ^k	20.52±0.34	-
τ^l	-	22.34±0.70
Unconst.	21.23±0.72	22.11±0.37
CVX	22.18±0.79	-
Dual (Ours)	**32.48±1.81**	**27.59±1.23**

4.1. Input perturbation

Our first task is to measure the explanation that is provided by our consistent attention compared to explanation baselines when learning to classify altered biodistribution levels in FDG-PET/CT patient scans. The biodistribution is categorically labeled to normal, mild, and severe and depicts the diffusion of FDG radiotracer, which serves as a marker for the tissue uptake of glucose, which in turn is closely correlated with certain types of tissue metabolism. The explanation in this experiment directs the interpreting physician to areas in the image with normal, mildly, and severely altered muscular FDG uptake.

Dataset: The ^{18}F-Fluorodeoxyglucose Positron Emission Tomography Muscular Uptake Altered Biodistribution (FDG-PET-MUAB) dataset contains 2637 images of 2470 patients that underwent a FDG-PET/CT scan in a nuclear medicine institute of a central health care campus during 2019 for various indications. Every image is a processed 2D coronal view of a maximum intensity projection (MIP) produced from a 3D PET DICOM file. For additional information regarding data collection, anonymization and annotation, please refer to the appendix.

Experimental setup: A baseline ResNet-50 (He et al., 2016) model was trained on a randomly sampled train set (80/20, patient disjoint) and obtained a balanced accuracy score of 83.6% on the test set. We used the balanced accuracy score as mildly and severely altered biodstiribution cases are typically infrequent. Next, we fine-tuned four variants of our method: τ^k with a single attention unit at the end of block 3; τ^l with a single attention unit at the end

of block 4; *Unconstrained* with two attention units (τ^k, τ^l) without a consistency term; *Dual* with two attention units and a consistency loss term (see Fig. 2).

Each attention unit is comprised of two cascaded potential gates (Eq. (3)). The attention units output two spatial distributions ($\tau^k \in \mathbb{R}^{32\times32}, \tau^l \in \mathbb{R}^{16\times16}$) highlighting salient regions in the input image. In addition, using τ^k, τ^l from the unconstrained variant, we solve Eq. (4) using a convex solver (Bynum et al., 2021) for each sample in the test set and output μ^k, μ^l as attribution maps (referred to as CVX model). The dual decomposition model was trained using the loss from Eq. (6), setting $\alpha = 10$ to match the average magnitude of loss terms. The rest of the models were trained using categorical cross-entropy loss. Adam (Kingma & Ba, 2014) optimizer was used, with a learning rate of 10^{-3} for all models. ViT-B/16 was fine-tuned using default hyper-parameters (12 attention heads and no dropout). Gradient and attribution propagation methods were computed with respect to the baseline model. The results of all fine-tuned models were averaged for five runs.

Following the evaluation scheme in (Samek et al., 2016; Feng et al., 2018; Chefer et al., 2020), given an input x and an attribution map, we rank the map elements by ascending importance. Subsequently, we perturb the input x by removing a growing amount of pixels according to the ranked attribution. We compare our method to the baselines by feeding the perturbed input back into the baseline ResNet-50 model and measuring the balanced accuracy score.

Results: Fig. (3) shows the input perturbation curve of our proposed *Dual* approach compared to explanations from attention variants and additional baselines. Tab. (1) shows the area under the curve for extended input perturbation curves from Fig. (3). Fig. (4) illustrates visual explanations on a FDG-PET-MUAB dataset sample, obtained by our method and the baselines. Additional illustration of visual explanations can be found in the appendix.

4.2. Implicit segmentation

Our second task is to measure the explanation obtained by our consistent attention compared to explanation baselines in magnetic resonance imaging (MRI). Given an axial 2D MRI slice of a brain hemisphere (left or right), the model learns to predict whether the hemisphere has a tumor or not. The explanation in this experiment directs the interpreting physician to the tumor's spatial location.

Dataset: The Brain Tumor Segmentation Challenge (BraTS18) dataset (Menze et al., 2014; Bakas et al., 2017; 2018) is a public dataset containing 285 MRI scans depicting pre-operative low and high-grade glioblastoma (LGG/HGG). Every sample includes four image modalities and a manually annotated segmentation mask. To obtain non-localized classification labels, we perform a vertical crop on every axial MRI slice, separating the tumor from the rest of the brain. As glioblastomas are usually common in one hemisphere, this produced two images from a given slice, each roughly with one hemisphere and a "yes" or "no" tumor label. We omitted the segmentation masks during training and included only binary labels as supervision. We discarded slices with low tumor area (less than 1000 pixels) and performed a random 80/20 train-test split (patient disjoint).

Experimental setup: A baseline CNN using a U-Net encoder (Ronneberger et al., 2015) was trained in a binary classification setting and achieved an accuracy score of 97.17% on the test set. Similarly to Section 4.1, we fine-tuned four variants of our attention units at two mid and high-level convolutional blocks of the network, outputting up to two spatial distributions ($\tau^k \in \mathbb{R}^{64\times64}$, $\tau^l \in \mathbb{R}^{32\times32}$). In addition to Section 4.1, we trained two variants in which we replaced the second loss term in Eq. 6 with a non-decomposed KL term. This term directly enforces consistency between the two attention blocks and is equivalent to setting $\lambda = 0$. Baselines and hyper-parameters are identical to the ones described in Section 4.1. We compare our method to the baselines by measuring the overlap between the scaled attribution and segmentation mask annotations. As attribution maps output "soft" boundaries to areas of interest, we threshold the attribution maps by three quantiles $(0.975, 0.95, 0.9)$ and measure the intersection over union (IoU) between the thresholded attribution and the segmentation mask. Furthermore, we compute the mean average precision score (mAP) between the attribution and segmentation masks, which is a threshold-agnostic metric. Our code for this experiment is publicly available[1].

Results: Tab. (2) summarizes the mAP and IoU metrics for explanations obtained from our proposed *Dual* approach compared to explanations from attention variants and additional baselines. Fig. (5) illustrates visual explanations on a BraTS18 dataset sample obtained by our method and the baselines. An additional illustration of visual explanations can be found in the appendix.

5. Conclusion and future work

Deploying machine learning algorithms in high-risk environments requires explaining their decisions to the user to ensure the model relies on the relevant anomalies for its decision, e.g., sites with abnormal FDG uptake in PET scans. Attention probability distributions are appealing as they both restrict the information flow through the neural network as well as visualize these restrictions as explanations. While setting multiple attention gates may improve the deep net's performance, it introduces inconsistencies between them.

[1] https://github.com/tomron27/dd_med

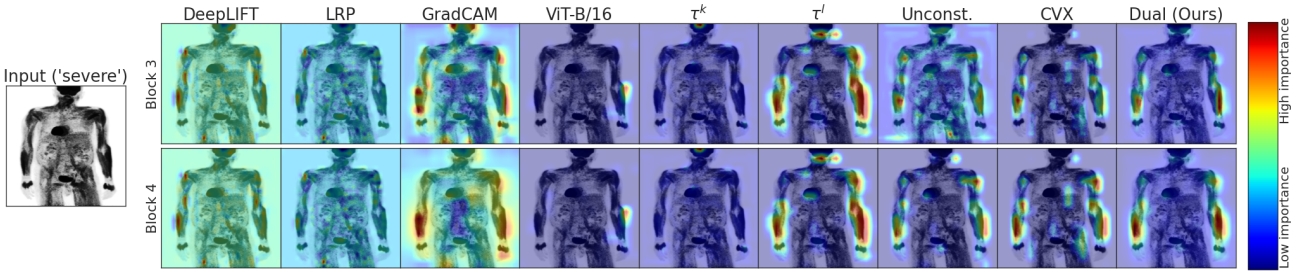

Figure 4. Explanations on a FDG-PET-MUAB dataset sample. The input image (leftmost) is labeled with severely altered biodistribution, which is evident mainly due to diffused uptake of FDG in both forearms, biceps, and deltoids. Compared to the baselines, we illustrate that our *Dual* approach (rightmost) assigns both high importance (in red) to target regions and near-zero importance (in blue) to other regions, thus providing a comprehensive and faithful explanation. Visualizations of single-level attribution maps (τ^k, τ^l) are duplicated across the opposite block.

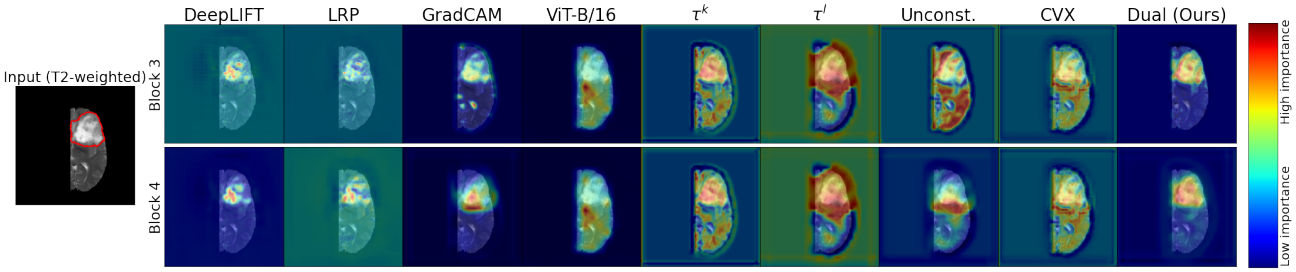

Figure 5. Brain tumor explanations on a vertically cropped BraTS18 dataset sample (positive instance). Input image and ground truth segmentation is in red contour (leftmost). Our *Dual* approach (rightmost) obtains the highest overlap with the ground-truth segmentation mask (which was not provided as supervision). Other attribution methods produce partial or noisy explanations. Visualizations of single-level attribution maps (τ^k, τ^l) are duplicated across the opposite block.

Thus, their explanatory value decreases and hinders the interpreting physician's ability to rely on the algorithm's decision. In this work, we suggest enforcing consistency using convex optimization. To avoid the computational complexity accompanying such operations, we further use a duality theorem to decompose the task into two reparameterized attention probability distributions. These reparameterizations allow us to learn consistent attention units without solving the primal convex optimization program.

There are a few limitations to our work. First, the choice of which convolutional blocks to apply our attention gates to was not exhaustively explored, as it is exponential in the network's depth. In addition, our proposed method outputs explanations that typically have a lower resolution compared to the input. Lastly, attention mechanisms are label-agnostic by nature, i.e., it is unclear what is the explanatory value of attention weight distribution when the predicted label indicates no findings in the medical image.

This work can be extended in several directions. With the success of deep nets, explainability in high-risk environments brings forth design choices for the learning algorithms. Consistent attention is such a choice that fits medical images; however, other design choices are required for different tasks and different data modalities. Also, learning convex optimization layers provides useful inductive bias, but they are slow to adopt. Dual decomposition can be used as a leading principle to improve their adoption beyond learning consistent probability models. Such general-purpose solvers require learning complex dual witnesses and are subject to further research.

Table 2. mAP and IoU results for implicit segmentation on the BraTS18 MRI dataset. We collect visual explanations (attributions) from our *Dual* method and multiple baseline explainers in two convolutional blocks (3,4). We measure the overlap between the explanation and ground truth segmentation at different spatial resolutions. Overlap was computed against positive halves of the source image (which contain a tumor). We compare our *Dual* approach against the baselines.

	Block 3			
Method	mAP	$IoU_{0.975}$	$IoU_{0.95}$	$IoU_{0.90}$
DeepLIFT	49.40	35.18	29.20	18.78
LRP	35.61	25.12	22.03	15.18
GradCAM	58.68	31.59	39.34	34.20
ViT-B/16	52.40±4.1	26.63±3.2	34.28±2.5	32.95±0.9
τ^k	33.03±11.4	18.19±8.1	25.79±7.1	23.25±4.6
Unconst.	19.01±12.3	9.89±8.5	13.82±11.1	14.32±+8.4
CVX	34.38±14.0	19.94±8.6	24.66±10.5	23.23±6.12
Dual (Ours)	**71.82±4.2**	**45.23±2.0**	**48.59±2.9**	32.75±2.3
	Block 4			
DeepLIFT	55.66	40.24	33.38	20.93
LRP	39.58	27.51	23.51	16.10
GradCAM	45.04	23.78	31.87	29.41
τ^l	34.17±9.0	17.88±4.9	24.20±5.2	23.89±4.2
Unconst.	41.44±10.0	23.65±5.5	29.19±5.6	26.09±4.0
KL($\tau^l \|\| \sum_{j\in N(i)} \tau^k$)	45.78±7.21	-	-	-
KL($\tau^k \|\| \sum_{j\in N(i)} \tau^l$)	45.56±7.14	-	-	-
Dual (Ours)	**76.16±3.8**	**46.17±1.7**	**51.29±2.7**	**34.64±2.0**

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., and Kolter, J. Z. Differentiable convex optimization layers. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/9ce3c52fc54362e22053399d3181c638-Paper.pdf.

Amos, B. and Kolter, J. Z. OptNet: Differentiable optimization as a layer in neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/amos17a.html.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., and Davatzikos, C. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K. W.-D., and McWilliams, B. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pp. 342–350. PMLR, 2017.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3286–3295, 2019.

Ben-Haim, S. and Ell, P. 18f-fdg pet and pet/ct in the evaluation of cancer treatment response. *Journal of Nuclear Medicine*, 50(1):88–99, 2009.

Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.-P., and Bach, F. Learning with differentiable pertubed optimizers. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*,

volume 33, pp. 9508–9519. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/6bb56208f672af0dd65451f869fedfd9-Paper.pdf.

Bertsekas, D. P., Nedic, A., and Ozdaglar, A. E. *Convex analysis and optimization*. Belmont, MA: Athena Scientific, 2003. ISBN 1-886529-45-0.

Bynum, M. L., Hackebeil, G. A., Hart, W. E., Laird, C. D., Nicholson, B. L., Siirola, J. D., Watson, J.-P., and Woodruff, D. L. *Pyomo–optimization modeling in python*, volume 67. Springer Science & Business Media, third edition, 2021.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. *arXiv preprint arXiv:2012.09838*, 2020.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 883–892. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/chen18j.html.

Cohade, C. Altered biodistribution on fdg-pet with emphasis on brown fat and insulin effect. In *Seminars in nuclear medicine*, volume 40, pp. 283–293. Elsevier, 2010.

Connor, C. E., Egeth, H. E., and Yantis, S. Visual attention: bottom-up versus top-down. *Current biology*, 14(19): R850–R852, 2004.

Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Domke, J. Implicit differentiation by perturbation. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/6ecbdd6ec859d284dc13885a37ce8d81-Paper.pdf.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., and Boyd-Graber, J. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*, 2018.

Fong, R., Patrick, M., and Vedaldi, A. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2950–2958, 2019.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

Gan, Z., Cheng, Y., Kholy, A., Li, L., Liu, J., and Gao, J. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6463–6474, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1648. URL https://www.aclweb.org/anthology/P19-1648.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, H., Zhang, Z., Xie, Z., and Lin, S. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3464–3473, 2019.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Jain, S. and Wallace, B. C. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Jain, S., Wiegreffe, S., Pinter, Y., and Wallace, B. C. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*, 2020.

Jetley, S., Lord, N. A., Lee, N., and Torr, P. H. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Zhao, T. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O. Captum: A unified and generic model interpretability library for pytorch, 2020.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Mahendran, A. and Vedaldi, A. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

Mangasarian, O. and Fromovitz, S. The fritz john necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17(1):37–47, 1967. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(67)90163-1. URL https://www.sciencedirect.com/science/article/pii/0022247X67901631.

Mehta, S., Ghazvininejad, M., Iyer, S., Zettlemoyer, L., and Hajishirzi, H. Delight: Very deep and light-weight transformer. *arXiv preprint arXiv:2008.00623*, 2020.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Nakajo, M., Jinnouchi, S., Fukukura, Y., Tanabe, H., Tateno, R., and Nakajo, M. The efficacy of whole-body fdg-pet or pet/ct for autoimmune pancreatitis and associated extrapancreatic autoimmune lesions. *European journal of nuclear medicine and molecular imaging*, 34(12):2088–2095, 2007.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Paulus, M. B., Choi, D., Tarlow, D., Krause, A., and Maddison, C. J. Gradient estimation with stochastic softmax tricks. *arXiv preprint arXiv:2006.08063*, 2020.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Rush, A. M. and Collins, M. A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing. *Journal of Artificial Intelligence Research*, 45:305–362, 2012.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. doi: 10.1109/JPROC.2021.3060483.

Schroeder, W., Martin, K., and Lorensen, B. *The Visualization Toolkit–An Object-Oriented Approach To 3D Graphics*. Kitware, Inc., fourth edition, 2006.

Schwartz, I., Schwing, A., and Hazan, T. High-order attention models for visual question answering. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/051928341be67dcba03f0e04104d9047-Paper.pdf.

Schwartz, I., Schwing, A. G., and Hazan, T. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12548–12558, 2019a.

Schwartz, I., Yu, S., Hazan, T., and Schwing, A. G. Factor graph attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2039–2048, 2019b.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Shammas, A., Lim, R., and Charron, M. Pediatric fdg pet/ct: physiologic uptake, normal variants, and benign conditions. *Radiographics*, 29(5):1467–1486, 2009.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/shrikumar17a.html.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Sra, S., Nowozin, S., and Wright, S. J. *Optimization for machine learning*. Mit Press, 2012.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/80537a945c7aaa788ccfcdf1b99b5d8f-Paper.pdf.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Wang, D., Gong, C., and Liu, Q. Improving neural language modeling via adversarial training. In *International Conference on Machine Learning*, pp. 6555–6565. PMLR, 2019.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2017.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020.

Yeung, H. W., Grewal, R. K., Gonen, M., Schöder, H., and Larson, S. M. Patterns of 18f-fdg uptake in adipose tissue and muscle: a potential source of false-positives for pet. *Journal of Nuclear Medicine*, 44(11):1789–1796, 2003.

Yu, D., Fu, J., Mei, T., and Rui, Y. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4709–4717, 2017.

Yu, M., Chang, S., Zhang, Y., and Jaakkola, T. S. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*, 2019.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., and Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 267–283, 2018.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

A. Appendix

B. Section 3.2: Theorem 1 Proof

Recall our primal program

$$\arg \min_{\mu^k, \mu^l \in \Delta} D_{KL}(\mu^k || \tau^k) + D_{KL}(\mu^l || \tau^l) \;\; \forall i \;\; \mu_i^l = \sum_{j \in N(i)} \mu_j^k \tag{7}$$

where Δ denotes a probability simplex.

Theorem B.1. *The probability distributions μ^k, μ^l are the optimal solutions to the primal program if*

$$\forall i \quad \mu_i^l = \frac{\tau_i^l e^{\lambda_i}}{\sum_s \tau_s^l e^{\lambda_s}} \tag{8}$$

$$\forall j \in N(i) \quad \mu_j^k = \frac{\tau_j^k e^{-\lambda_i}}{\sum_t \sum_{s \in N(t)} \tau_s^k e^{-\lambda_t}} \tag{9}$$

$$\forall i \quad \mu_i^l = \sum_{j \in N(i)} \mu_j^k \tag{10}$$

Proof. Under probability simplex constraints, a corresponding Lagrangian of (7) is

$$L(\mu, \lambda) = D_{KL}(\mu^k || \tau^k) + D_{KL}(\mu^l || \tau^l) + \sum_{i=1}^{d_l} \lambda_i (\mu_i^l - \sum_{j \in N(i)} \mu_j^k). \tag{11}$$

From strong duality, we can assert that μ^*, λ^* are a primal-dual pair if the following three conditions hold (cf. Convex Analysis and Optimization (Bertsekas et al., 2003), Section 7.2):

1. $\mu^* = \arg\min_{\mu^k, \mu^l \in \Delta} L(\mu, \lambda^*)$
2. $\lambda^* = \arg\max_{\lambda} L(\mu^*, \lambda)$
3. $\mu^* \in \Delta$

Examining the first condition 1:

$$\arg \min_{\mu^k, \mu^l \in \Delta} L(\mu, \lambda^*) = \tag{12}$$

$$= \arg \min_{\mu^k, \mu^l \in \Delta} D_{KL}(\mu^k || \tau^k) + D_{KL}(\mu^l || \tau^l) + \sum_{i=1}^{d_l} \lambda_i^* (\mu_i^l - \sum_{j \in N(i)} \mu_j^k) \tag{13}$$

$$= \arg \min_{\mu^k, \mu^l \in \Delta} D_{KL}(\mu^k || \tau^k) + D_{KL}(\mu^l || \tau^l) + \sum_{i=1}^{d_l} \mu_i^l \lambda_i^* - \sum_{i=1}^{d_l} \sum_{j \in N(i)} \mu_j^k \lambda_i^* \tag{14}$$

$$= \arg \min_{\mu^k, \mu^l \in \Delta} \sum_{j=1}^{d_k} \mu_j^k log(\frac{\mu_j^k}{\tau_j^k}) + \sum_{i=1}^{d_l} \mu_i^l log(\frac{\mu_i^l}{\tau_i^l}) + \sum_{i=1}^{d_l} \mu_i^l \lambda_i^* - \sum_{i=1}^{d_l} \sum_{j \in N(i)} \mu_j^k \lambda_i^* \tag{15}$$

$$= \arg \min_{\mu^k, \mu^l \in \Delta} \sum_{i=1}^{d_l} \sum_{j \in N(i)} \mu_j^k (log(\frac{\mu_j^k}{\tau_j^k}) - \lambda_i^*) + \sum_{i=1}^{d_l} \mu_i^l (log(\frac{\mu_i^l}{\tau_i^l}) + \lambda_i^*) \tag{16}$$

$$= \arg \min_{\mu^k, \mu^l \in \Delta} \sum_{i=1}^{d_l} \sum_{j \in N(i)} \mu_j^k log(\frac{\mu_j^k}{\tau_j^k e^{\lambda_i^*}}) + \sum_{i=1}^{d_l} \mu_i^l log(\frac{\mu_i^l}{\tau_i^l e^{-\lambda_i^*}}) \tag{17}$$

$$= \arg \min_{\mu^k, \mu^l \in \Delta} D_{KL}(\mu^k || \tau^k \odot e^{\lambda^*}) + D_{KL}(\mu^l || \tau^l \cdot e^{-\lambda^*}). \tag{18}$$

For any i it holds that $(\tau^l \cdot e^{-\lambda^*})_i = \tau_i^l e^{-\lambda_i^*}$. For any i and $j \in N(i)$ it holds that $(\tau^k \odot e^{\lambda^*}) = \tau_j^k e^{\lambda_i^*}$. Since λ is unconstrained, the second condition 2 translates to $\frac{\partial}{\partial \lambda} L(\mu^*, \lambda^*) = 0$ which in turn enforces $\mu_i^{*l} = \sum_{j \in N(i)} \mu_j^{*k}$ for every i. The probability distributions minimizing (18) are the normalized arguments, namely

$$\forall i \quad \mu_i^{*l} = \frac{\tau_i^l e^{\lambda_i^*}}{\sum_s \tau_s^l e^{\lambda_s^*}} \tag{19}$$

$$\forall j \in N(i) \quad \mu_j^{*k} = \frac{\tau_j^k e^{-\lambda_i^*}}{\sum_t \sum_{s \in N(t)} \tau_s^k e^{-\lambda_t^*}}. \tag{20}$$

□

C. Experiments

C.1. Section 4.1: Input perturbation on FDG-PET-MUAB dataset

C.1.1. FDG-PET/CT

^{18}F-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography (^{18}F-FDG PET/CT) is a medical imaging technique widely used in oncology and cancer staging. A PET scan involves injecting the patient with a radioactive glucose analog (^{18}F-FDG), which binds to tissue with high metabolic activity. The radiation emitted by FDG is then traced back to its origin, and sites with high glucose uptake (such as lesions) are visualized. As PET scans typically output volumetric images, Maximal Intensity Projection (MIP) is applied to obtain 2D images onto several viewing planes. MIP rendering is a standard visualization technique in nuclear medicine, and it serves as a principal viewpoint in diagnosis by radiologists and physicians (Ben-Haim & Ell, 2009; Shammas et al., 2009; Nakajo et al., 2007).

One of the most critical limitations of PET is that FDG accumulates in normal tissues with high metabolic activity, such as the brain and muscles. Altered biodistribution (Cohade, 2010; Yeung et al., 2003) is a state in which FDG abnormally accumulates in benign tissue and might render PET images uninterpretable to medical physicians by obscuring nearby lesions or depleting FDG reserves.

C.1.2. FDG-PET-MUAB Dataset Acquisition

The ^{18}F-Fluorodeoxyglucose Positron Emission Tomography Muscular Uptake Altered Biodistribution (FDG-PET-MUAB) dataset contains 2637 images of 2470 patients that underwent a ^{18}F-FDG PET/CT scan in a nuclear medicine institute of a central health care campus during 2019 for various indications. Every image is a processed 2D MIP (coronal view) of an original WB 3D MAC PET DICOM file. The source DICOM files were acquired by a GE Discovery 690 PET/CT scanner. Each axial slice of the 3D DICOM has a physical thickness of 3.27mm and was resampled onto a spatial dimension of 256×256 pixels. The 3D PET image is a $256 \times 256 \times n_i$ sized tensor, where n_i varies among scans. Pixel values are in 16bit and are normalized to a SUV_{bw} standard by: $SUV_{bw}(x_i) = (x_i * w_i)/d_i$ where x_i is the pixel value, w_i is the patient weight at the time of the scan (in grams), and d_i is the FDG dosage (in bqml). The coronal MIP image of the 3D tensor is obtained by a ray casting mapper implemented by the VTK (Schroeder et al., 2006) library. A further histogram equalization (256 bins) is applied on the flattened image, and a final rescaling to 512×512 pixels is performed. All personally identifiable information of patients was discarded during the preparation of the dataset. Coronal MIP images contain a single view of a diffused bio-chemical distribution and are not anatomically detailed, hence cannot be personally identifiable.

A trained graduate student, overseen by a nuclear medicine expert, annotated scans for three levels of altered biodistribution severity: normal, mild, and severe. Annotation was performed using a self developed 3D PET/CT tagging tool based on the VTK (Schroeder et al., 2006) library. The tagging tool can display MIP from multiple viewpoints, allowing higher confidence in annotation. The dataset is highly imbalanced as only roughly 20% of the samples are labeled as "mild" and roughly 3% are of are labeled as "severe". We consider the FDG-PET-MUAB dataset to be a weakly annotated dataset, as it does not contain bounding boxes or segmentation masks that localize sites with abnormal FDG uptake.

C.2. Section 4.2: Implicit segmentation on BraTS18 dataset

Our code for this experiment is publicly available at: `https://github.com/tomron27/dd_med`.

Table 3. Accuracy metrics and parameters counts for attention models in FDG-PET-MUAB and BraTS18 datasets.

Model		Baseline	τ^k	τ^l	Unconst.	Dual
ResNet-50 (FDG-PET-MUAB	B. Acc.	83.66	82.56±0.67	80.82±0.25	83.49±0.66	82.79±0.46
	# Param.	23.5M	26.6M	27.7M	30.8M	30.8M
ViT-B/16 (FDG-PET-MUAB)	B. Acc.	74.99±1.67	-	-	-	-
	# Param.	85.4M	-	-	-	-
U-Net Encoder (BraTS18)	Acc.	97.17	97.09±0.25	97.30±0.19	97.26±0.17	97.36±0.26
	# Param.	4.7M	4.7M	4.9M	4.9M	4.9M
ViT-B/16 (BraTS18)	Acc.	97.31±0.44	-	-	-	-
	# Param.	85.4M	-	-	-	-

C.3. FDG-PET-MUAB and BraTS18 visual examples