Targeted Activation Penalties Help CNNs Ignore Spurious Signals

Dekai Zhang[1], **Matt Williams**[2,3], **Francesca Toni**[1]

[1]Department of Computing, Imperial College London
[2]Department of Radiotherapy, Charing Cross Hospital
[3]Institute of Global Health Innovation, Imperial College London
{dz819, matthew.williams, f.toni}@imperial.ac.uk

Abstract

Neural networks (NNs) can learn to rely on spurious signals in the training data, leading to poor generalisation. Recent methods tackle this problem by training NNs with additional ground-truth annotations of such signals. These methods may, however, let spurious signals re-emerge in deep convolutional NNs (CNNs). We propose ***T**argeted **A**ctivation **P**enalty (TAP)*, a new method tackling the same problem by penalising activations to control the re-emergence of spurious signals in deep CNNs, while also lowering training times and memory usage. In addition, ground-truth annotations can be expensive to obtain. We show that TAP still works well with annotations generated by pre-trained models as effective substitutes of ground-truth annotations. We demonstrate the power of TAP against two state-of-the-art baselines on the MNIST benchmark and on two clinical image datasets, using four different CNN architectures.

Introduction

Neural networks (NNs) have demonstrated strong performances and in some domains have exceeded experts (Rajpurkar et al. 2018; Ke et al. 2021). These success stories come with an important caveat: NNs appear to be very good at exploiting spurious signals which boost their performance on the training data but lead to poor generalisation (Hendricks et al. 2018). Ribeiro, Singh, and Guestrin (2016), for instance, find that a model can achieve near perfect results in distinguishing huskies from wolves by making use of the background in the images. More worryingly, some models trained to detect pneumonia in chest radiographs (Zech et al. 2018) were found to make heavy use of image artifacts.

Recent work on *explanatory supervision (XS)* has shown that models can be successfully protected from learning spurious signals by eliciting ground-truth explanations from humans as an additional supervision for the models (Ross, Hughes, and Doshi-Velez 2017; Teso and Kersting 2019; Rieger et al. 2020; Schramowski et al. 2020; Hagos, Curran, and Mac Namee 2022; Friedrich et al. 2023). Of these methods, *"right for the right reasons" (RRR)* (Ross, Hughes, and Doshi-Velez 2017) and *"right for better reasons" (RBR)* (Shao et al. 2021) seem to perform particularly well (Friedrich et al. 2023). These convincing performances, however, have been obtained on datasets requiring relatively shallow models (i.e., VGG-16 (Simonyan and Zisserman 2015) and shallower). It appears that they may be less successful with deeper convolutional NNs (CNNs), as illustrated under "RRR" and "RBR" in Figure 1. Also, these works assume that extensive ground-truth explanations can be obtained, typically from humans, which in practice can be costly, especially in expert domains such as healthcare.

In this paper, we define a novel XS method for CNNs, which adds a ***T**argeted **A**ctivation **P**enalty (TAP)* to spurious signals, mitigating against the re-emergence of spurious signals in deeper layers (illustrated under "TAP" in Figure 1). We show that TAP performs competitively and better than RRR and RBR, while requiring lower training times and memory usage. We further show that TAP can still perform well when replacing ground-truth explanations with noisier annotations generated by a teacher model pre-trained on as little as 1% of the target domain. Our contributions can be summarised as follows[1]:

- We introduce TAP to teach CNNs to ignore spurious signals. We formally relate TAP to RRR and RBR: whereas RRR and RBR directly target input gradients, TAP indirectly does so by minimising activations, avoiding expensive second-order derivatives. In our experiments, this results in circa 25% of the training time and half the memory usage compared to RRR and RBR.
- We compare TAP to RRR and RBR in the standard setting of a human teacher who provides ground-truth annotations, and we demonstrate that TAP can still be effective with noisy but automatically generated explanations from a teacher model which is pre-trained on a small clean dataset.
- Our findings are supported by experiments (i) on MNIST (LeCun, Cortes, and Burges 1998) using a simple two-layer CNN as a standard benchmark and, to show the efficacy of TAP on higher-stakes real world datasets, (ii) on two clinical datasets for pneumonia (Kermany et al. 2018) and osteoarthritis (Chen et al. 2019) using three commonly used architectures: VGG-16, ResNet-18 (He et al. 2016) and DenseNet-121 (Huang et al. 2017).

The extended version of this paper (with Appendix) is deposited at: https://arxiv.org/abs/2311.12813

[1]Source code: https://github.com/dkaizhang/TAP

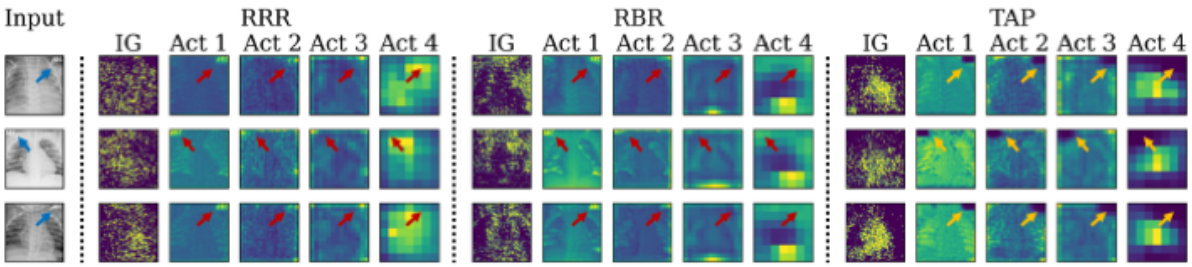

Figure 1: Three chest x-rays with spurious tags placed in the corners (blue arrows). Input gradients (IG) and activations at each of the four convolutional blocks (Act 1–4) of a ResNet-18 trained using RRR, RBR and TAP. With RRR and RBR the spurious tags can re-emerge in deeper layers (red arrows), while TAP mitigates their presence throughout (yellow arrows).

Related Work

Explainable AI (XAI). NNs are frequently seen as black boxes which can be difficult to audit (Adler et al. 2018). In computer vision, saliency maps are often used to highlight relevant areas in a given input (Simonyan, Vedaldi, and Zisserman 2014; Springenberg et al. 2015; Sundararajan, Taly, and Yan 2017; Arrieta et al. 2019; Selvaraju et al. 2020) and have been successfully deployed to identify models which use spurious signals in the data (Ribeiro, Singh, and Guestrin 2016; Lapuschkin et al. 2019). While XAI offers tools of identification, it typically does not address how to correct models that use spurious signals.

Explanatory Supervision (XS). Recent work on XS has investigated if ground-truth annotations of spurious signals can be used as an additional source of supervision to train models to make predictions for the "right reasons" (Ross, Hughes, and Doshi-Velez 2017; Teso and Kersting 2019; Rieger et al. 2020; Schramowski et al. 2020; Shao et al. 2021; Heo et al. 2023). The pioneering method, RRR, by Ross, Hughes, and Doshi-Velez (2017) and RBR, a later extension of RRR by Shao et al. (2021), use input gradient regularisation, which have been shown to be an effective mechanism to prevent models from learning spurious relationships (Friedrich et al. 2023). These, however, rely on expensive second-order derivatives, whereas our proposed method does not. Furthermore, common to the above methods is the assumption of sample-wise ground-truth annotations, typically provided by a human. We find that, with our novel method, ground-truth annotations can in some cases be replaced with noisier annotations from a pre-trained teacher model.

Teacher-Student Settings. The setting we consider is closest to knowledge distillation (Hinton, Vinyals, and Dean 2015) in which a student model receives supervision from a teacher model. Specifically, the student is trained to match logit targets provided by the teacher. Later extensions also match activations (Zagoruyko and Komodakis 2017) or Jacobians (Srinivas and Fleuret 2018). The setting we consider differs in that the teacher does not directly provide additional targets to be matched but instead identifies areas to be ignored. We discuss these methods further in the extended paper.

Preliminaries

Setting. Assume a *dataset* $\{(X_i, y_i)\}_{i=1}^N$ of N labelled images. For brevity, we assume that each image is single-channel, so that $X_i \in \mathcal{X} \subseteq \mathbb{R}^{H,W}$ consists of $H \times W$ pixels. Each label $y_i \in \mathcal{Y} \subseteq \{0,1\}^K$ is a one-hot vector over K classes. Suppose we wish to learn, from this dataset, a CNN with L convolutional layers, given by f_θ, parameterised by θ, such that for every *input* $X \in \mathcal{X}$, $f_\theta(X) = \hat{y}$ with $\hat{y} \in \{0,1\}^K$ denoting the *output vector*.

Similar to Adebayo et al. (2022), we focus on a setting where the dataset is the result of a label-revealing *contamination* of a *clean* dataset $\{(X_i^*, y_i)\}_{i=1}^N$ where $X_i^* \in \mathcal{X}^* \subseteq \mathbb{R}^{H,W}$. The contaminated instance X_i thus contains a signal which spuriously reveals y_i, whereas X_i^* does not. Formally, we suppose there exists some *spurious contamination function* $SC : \mathcal{X}^* \times \mathcal{Y} \to \mathcal{X}$ which induces spurious correlations in the clean dataset. Examples of spurious signals include medical tags, which are text strings frequently found on radiographs and which may spuriously correlate with the label (Adebayo et al. 2022). The objective in this setting is to learn a classifier f_θ from the contaminated data that does not only classify well but which does so without relying on spurious signals. Intuitively, the classifier should produce the same output regardless of what spurious signals are added by the contamination function.

Definition 1. *Classifier f_θ is* not reliant on spurious signals $SC(X^*, y) \in \mathcal{X}$ *for* $X^* \in \mathcal{X}^*, y \in \mathcal{Y}$ *if:*

$$f_\theta(SC(X^*, y)) = f_\theta(SC(X^*, permute(y))) \quad (1)$$

where $permute(y)$ is some permutation of label y.

Learning to Ignore Spurious Signals. If the model is making correct classifications but is found to use spurious signals, it is not "right for the right reasons" (Ross, Hughes, and Doshi-Velez 2017). To revise the model, the common approach in XS is to obtain feedback, typically from a human, in the form of an annotation mask $M \in \mathbb{R}^{H,W}$ for each image, which indicates spurious signals therein.

To construct these masks, observe that each input $X \in \mathcal{X}$ can be decomposed into two disjoint matrices holding *clean signals* X^c and *spurious signals* X^s, so that $X = X^c + X^s$. Note that X^c therefore only contains the parts of the clean input X^* unaffected by the transformation SC, so that, for

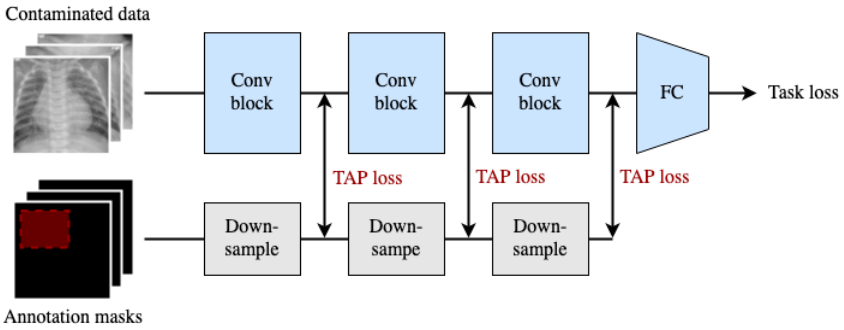

Figure 2: TAP losses target activations throughout the CNN with (input-level) annotation masks.

$i = 1, \ldots, H, j = 1, \ldots, W$, $X^c_{ij} = X^*_{ij}$ if and only if $X^c_{ij} \neq 0$. We then define the mask as an indicator of spurious signals:

$$M_{ij} = \mathbb{1}(X^c_{ij} = 0) \tag{2}$$

Given mask M, a common approach in XS is to augment the loss function $\mathcal{L}$ to supervise both the task and reasons:

$$\mathcal{L}(X, y, M; \theta) = \mathcal{L}_{Task}(X, y; \theta) + \lambda \mathcal{L}_{XS}(X, M; \theta) \tag{3}$$

where $\mathcal{L}_{Task}$ corresponds to the task (e.g., classification) loss, $\mathcal{L}_{XS}$ is the XS loss penalising the use of wrong reasons as defined by M, and λ is a hyperparameter for the relative impact. Note the masks are not used to cover the input but define a penalty region, so that the model still "sees" the entire input.

RRR is a pioneering instantiation of $\mathcal{L}_{XS}$, which targets spurious signals by penalising their input gradients (Ross, Hughes, and Doshi-Velez 2017):

$$\mathcal{L}^{RRR}_{XS}(X, M; \theta) = \sum_{i=1}^{H} \sum_{j=1}^{W} \left[M \odot \sum_{k=1}^{K} \frac{\partial f_\theta(X)_k}{\partial X} \right]^2_{ij} \tag{4}$$

where $\odot$ denotes the Hadamard product. RBR, an extension of RRR, proposes to multiply influence functions with input gradients to take into account model changes from perturbing the input (Shao et al. 2021). Other instantiations exist (Rieger et al. 2020; Schramowski et al. 2020) but RRR and RBR appear to have the greatest corrective effect amongst XS methods (Friedrich et al. 2023). These methods, however, require higher-order derivatives which are expensive to compute. In this paper, we propose targeting activations of CNNs, which are computed as part of a single forward pass.

Our Approach

We introduce TAP, a new XS loss which targets activations of spurious signals with penalties (as illustrated in Figure 2). We then frame the objective of teaching a model to ignore spurious signals as a teacher-student problem in which the teacher is a pre-trained model (see Figure 3).

TAP: Targeted Activation Penalty

TAP takes advantage of the fact that convolutional layers preserve spatial relationships within an image in its activation map (LeCun, Bengio, and Hinton 2015). Given this, we propose targeting parts of the activation map corresponding to spurious signals with penalties.

Suppose $Z^l \in \mathbb{R}^{C^l, H^l, W^l}$ is the tensor output of the l-th convolutional layer with C^l output channels, filters w^l, bias b^l and activation function σ, so that:

$$Z^l = w^l * \sigma(Z^{l-1}) + b^l \tag{5}$$

where the $*$-operator denotes the convolutional product. We define the *activation map* A^l as the channel-wise sum:

$$A^l = \sum_{c=1}^{C^l} \sigma(Z^l_c) \tag{6}$$

Given an image $X \in \mathbb{R}^{H,W}$, where $H \geq H^l$ and $W \geq W^l$, we may need to downscale the annotation mask $M \in \mathbb{R}^{H,W}$ to match the dimensions of the activation map. We define a *downscaling function* $\mathcal{D} : \mathbb{R}^{H,W} \times \mathbb{N} \to \mathbb{R}^{H^l, W^l}$, which accepts the input-level annotation mask M and an integer to identify the layer and outputs a downscaled annotation mask M^l with the target dimensions. To effectively target the areas of the spurious signals in the activations, the downscaling function needs to preserve the spatial information in the annotation masks and account for the receptive field of a given activation element. We apply max-pooling with a stride of 1 using a kernel size of κ to effectively increase the annotation region by $\lfloor \kappa/2 \rfloor$ pixels in each direction. We then apply average pooling to match the height and width dimensions of layer l:

$$\mathcal{D}(M, l) = \text{AvgPool}(\text{MaxPool}(M), H^l, W^l) \tag{7}$$

We found a choice of $\kappa = 3$ to be a good balance between capturing the influence of targeted pixels and preventing over-regularisation from too large a penalty region.

Definition 2. *Given a downscaling function $\mathcal{D}$, annotation mask M and an L-layer CNN parameterised by θ, producing activation maps A^l in layer l, the* Targeted Activation Penalty (TAP) *is:*

$$\mathcal{L}^{TAP}_{XS}(X, M; \theta) = \sum_{l=1}^{L} \frac{1}{L} \left\| \mathcal{D}(M, l) \odot A^l \right\| \tag{8}$$

where $\|\cdot\|$ denotes the L1-norm.

In practice, for deep CNNs, instead of targeting the output of every convolutional layer, we focus on a subset (as envisaged in Figure 2).

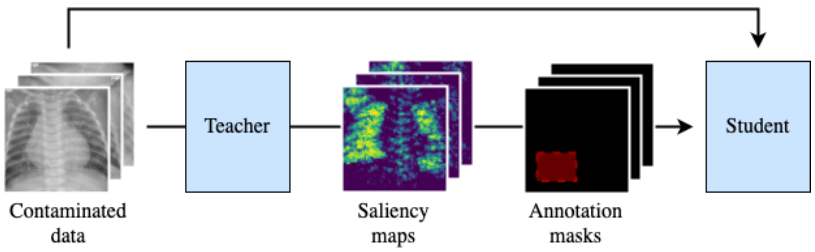

Figure 3: We use a pre-trained teacher model to derive saliency-based masks for a contaminated dataset to determine the least important regions in each input, resulting in an input-level mask for each input. The contaminated dataset and the corresponding masks are used to train a student model which learns to ignore the regions identified by the masks.

Relationship with RRR and RBR. Both RRR and RBR directly reduce the input gradient of spurious signals to zero. We show in Theorem 1 that for CNNs with ReLU activation function σ we can target elements in the activation map, as we do with TAP, to reduce the input gradient of pixel (i, j) for some convolutional layer $l > 1$, which is given by:

$$\frac{\partial}{\partial X_{ij}}\sigma(Z^l) = \sigma'(Z^l) \odot w^l * \frac{\partial \sigma(Z^{l-1})}{\partial X_{ij}} \tag{9}$$

Theorem 1. *Suppose we have a CNN with convolutional layers $l = 1, \ldots, L$, filters w^l of kernel size κ^l and ReLU activation σ. It is sufficient to optimise all activation map elements A^l_{ab} with respect to w^l to set the input gradient of pixel (i, j) in layer l (Equation 9) to zero, where (a, b) are such that:*

$$\begin{aligned} i - \kappa^1 - \ldots - \kappa^l + l \leq a \leq i \\ j - \kappa^1 - \ldots - \kappa^l + l \leq b \leq j \end{aligned} \tag{10}$$

We provide a sketch of the proof here and defer details to the extended paper. Note that activation elements (a, b) outside the area defined by Equation 10 do not capture pixel (i, j) within their receptive fields, so that the input gradient of (i, j) is zero for such elements. To ensure that it is zero for elements (a, b) inside the area, we optimise activation A^l_{ab} with respect to filters w^l_c, $c = 1, \ldots, C^l$. The first-order condition sets the following to zero:

$$\frac{\partial A^l_{ab}}{\partial w^l_c} = \sigma'(Z^l_{cab}) \begin{bmatrix} \delta^{l-1}_{a,b} & \cdots & \delta^{l-1}_{a+\kappa^l-1,b} \\ \vdots & \ddots & \vdots \\ \delta^{l-1}_{a,b+\kappa^l-1} & \cdots & \delta^{l-1}_{a+\kappa^l-1,b+\kappa^l-1} \end{bmatrix} \tag{11}$$

where $\delta^{l-1} = \sigma(Z^{l-1})$ which is constant with respect to w^l. Given ReLU activation, the optimisation results in $\sigma'(Z^l_{ab})$ being pushed to zero.

Figure 1 illustrates that TAP indeed mitigates the presence of spurious signals in the input gradients even when only a subset of the corresponding activations are targeted. We provide further quantitative evidence in the extended paper, showing that, with TAP, less than 1% of input gradients in the top-quartile by magnitude overlap with spurious signals. Notably, TAP does not require computing second-order derivatives to do so, which results in faster training times and lower memory consumption compared to RRR and RBR (Table 1).

	VGG-16		ResNet-18		Dense-121	
Loss	it/s	mem	it/s	mem	it/s	mem
No XS	12.1	4.4	38.5	2.5	11.6	4.4
RRR	2.6	9.7	12.5	3.2	3.0	9.2
RBR	1.7	8.6	7.3	4.0	OOM	
TAP	**11.8**	**4.4**	**36.7**	**2.5**	**12.0**	**4.5**

Table 1: Iterations per second (it/s) and memory usage in GB (mem) for batches of 16 224×224-pixel images. DenseNet-121 with RBR runs out of memory (OOM). TAP adds less training overhead. Best result amongst XS methods in bold.

A Teacher-Student Framework

Recognising spurious signals necessitates external input, as by definition they are informative in the training distribution. In previous works, the external input takes shape in the form of sample-wise ground-truth annotation masks $\{M_i\}_{i=1}^N$ (Ross, Hughes, and Doshi-Velez 2017; Rieger et al. 2020; Schramowski et al. 2020; Shao et al. 2021), which can be costly when obtained from humans. To address this issue, we propose framing the objective of teaching a model to ignore spurious signals as a teacher-student problem. We consider the teacher as a model parameterised by θ_T which transfers its knowledge to a student model parameterised by θ_S to be trained on contaminated data, so that:

$$\theta_S = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(X_i, y_i, \theta_T; \theta) \tag{12}$$

We instantiate this framework first with the standard setting in which the teacher model is a human who can provide ground-truth annotations and then with a pre-trained model which may provide noisier annotations. We argue that obtaining a small clean dataset $\{(X^*_i, y_i)\}_{i=1}^{N^*}$, with $N^* \ll N$, for pre-training a teacher may be easier than annotating all of the contaminated data. We propose using saliency maps to identify areas in the contaminated data the teacher model believes to be unimportant, which, if the teacher fulfils Equation 1, should include the spurious signals. To find these areas, we make use of an explanation function $\mathcal{E}$ which accepts an NN and an input to produce a saliency map over the input:

$$\mathcal{E}(X, y; \theta_T) = E \tag{13}$$

where $E \in \mathbb{R}^{H,W}$. In our experiments, we use input gradients as the explanation method, but the choice is not restricted to a particular saliency method. For notational convenience, we assume saliency maps E contain absolute values normalised to $[0, 1]$. We propose using an element-wise threshold function Ψ to construct the *teacher annotation* $\tilde{M}$:

$$\tilde{M}_{ij} = (\Psi \circ E)_{ij} = \begin{cases} 1 & \text{if } E_{ij} < \tau \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

for $i = 1, \ldots, H, j = 1, \ldots, W$. Intuitively, low saliency areas will be targeted by the teacher annotation. We can now re-state the optimal student parameters as:

$$\theta_S = \arg\min_{\theta} \sum_{i=0}^{N} \mathcal{L}(X_i, y_i, \tilde{M}_i; \theta) \tag{15}$$

Experimental Design

We first compare TAP against RRR and RBR with ground-truth annotations on a benchmark and two clinical datasets using four different CNN architectures. We then study performances under teacher annotations.

Datasets

We conduct experiments on MNIST, a standard benchmark in XS, and two clinical datasets: chest radiographs for detecting pneumonia (PNEU) and knee radiographs for grading osteoarthritis (KNEE). MNIST contains 60,000 28×28 pixel images of handwritten digits. PNEU contains 5,232 paediatric chest radiographs, with approximately two-thirds presenting pneumonia. KNEE contains 8,260 radiographs of knee joints with 5 different osteoarthritis grades on an ordinal scale. The images in PNEU and KNEE vary in resolution and dimensions. We centre-crop and resize to 224×224 pixels. We use the pre-defined training and test splits for all datasets and reserve 10% of the training split for validation.

To contaminate the datasets, we follow the common approach in XS and add spurious signals to the data which correlate with the label. Figure 4 shows samples of the contaminations. For MNIST, we follow Ross, Hughes, and Doshi-Velez (2017) and define $SC(X^*, y)$ (Definition 1) as adding 4×4 pixel patches to a random corner in X^* with a pixel value set to $255 - label \times 25$. For $SC(X^*, permute(y))$ (the *permuted* data), we choose random permutations of the labels, which essentially randomises the patch assignment. For the medical datasets we add text strings to simulate medical tags. For PNEU, $SC(X^*, y)$ adds "ABC" for normal and "XYZ" for pneumonic cases. Given the binary prediction task, we choose to swap instead of randomising labels to construct $SC(X^*, permute(y))$. For KNEE, $SC(X^*, y)$ adds "ABC" (grade 0), "DEF" (grade 1), "GHI" (grade 2), "JKL" (grade 3) and "MNO" (grade 5), while $SC(X^*, permute(y))$ randomises the string assignment. We consider experiments with an additional contamination of the medical datasets in the extended paper.

For the teacher-student experiments, we hold out 10% or 1% of the training split as a clean dataset to pre-train a teacher before adding artifacts to induce spurious correlations in the remainder, as described above.

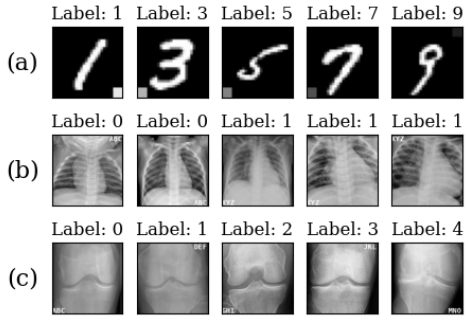

Figure 4: Overview of contaminated data. Brightness of corners correlate with label for (a) MNIST. Strings in corners correlate with label for (b) PNEU and (c) KNEE.

Evaluation Metrics

If a model exploits spurious signals, it should perform worse when the correlations are removed or changed, so that Equation 1 does not hold. To evaluate the models, we (i) assess their performance on the contaminated data $SC(X^*, y)_{test}$. We measure accuracy for MNIST, F-score for PNEU given class imbalance, and mean absolute error (MAE) for KNEE given ordinal scores. We then (ii) assess the sensitivity of that initial performance to spurious signals by measuring the change in the respective metrics (Δ-metric) when evaluated on $SC(X^*, permute(y))_{test}$. We report averaged results from 5 different seeds.

Models

We use a two-layer CNN (for architectural details, see the extended paper) for MNIST and three commonly used CNN architectures for the medical datasets: VGG-16, ResNet-18 and DenseNet-121. For the teacher-student experiments, we use ResNet-18 as teacher for the medical datasets and the two-layer CNN for MNIST.

We apply TAP to both layers of the two-layer CNN. For ResNet-18, we found the output of each of the four residual blocks to be a natural choice. For DenseNet-121, we analogously chose the output of each of the four dense blocks. For VGG-16, we choose the output after each of the MaxPool layers. In each case, the targeted layers are spaced roughly evenly across the depth of the architectures.

We choose cross-entropy for the task loss $\mathcal{L}_{Task}$. We use SGD as optimiser with weight decay of 0.9. We train for 50 epochs. We use random initialisation for the two-layer CNN and use a learning rate of 10^{-3}. For VGG-16, ResNet-18 and DenseNet-121 we initialise with ImageNet-weights and use a learning rate of 10^{-5}. We use a batch size of 256 for MNIST and 16 for the medical datasets (8 for DenseNet-121 with RBR, given memory constraints). We tune λ (Equation 3) by training on $SC(X^*, y)_{train}$ and evaluating the validation loss on $SC(X^*, permute(y))_{val}$. We increase λ in log steps from 10^{-9} to 1 for TAP and RRR and 1 to 10^9 for RBR (given much smaller losses)—see extended paper.

Experiments were implemented with PyTorch 1.13 and run on a Linux Ubuntu 18.04 machine with an Nvidia RTX 3080 GPU with 10GB VRAM.

Results

We show that TAP-trained models perform well on contaminated data and do not rely on spurious signals on the three chosen datasets. Below we report experiments on the common XS setting with ground-truth annotations. Throughout, we use as baselines "Base" models trained on clean data and "No XS" models trained on contaminated data without XS. We then consider the second setting with teacher annotations. In the extended paper, we expand on our experiments, e.g., (i) we analyse input gradients and activations quantitatively and visually, finding that TAP reduces the presence of spurious signals in deeper layers and, indirectly, in input gradients; (ii) we report results on another contamination of the medical datasets, further supporting our findings below; (iii) we report tabular values with standard deviations and (iv) include additional results on clean datasets; (v) we provide further evidence of TAP's efficacy on additional CNNs; (vi) we report an extended comparison of XS methods with teacher annotations; and, lastly, (vii) we discuss and compare with knowledge distillation methods.

Setting 1: Ground-Truth Annotations

MNIST. TAP performs competitively with existing methods. The left panel of Figure 5 compares the test performance on $SC(X^*, y)_{test}$ of different XS losses (RRR, RBR, TAP) against Base and No XS models. At first glance, the different losses result in very similar performances. The right panel shows the drop in accuracy when evaluating on $SC(X^*, permute(y))_{test}$. Models that rely on spurious signals should be sensitive to changes in the spurious correlations. As expected, the accuracy of the No XS model drops significantly, whereas the Base model and the models trained with XS losses are insensitive to the spurious signals. This demonstrates that TAP performs as well as RRR and RBR.

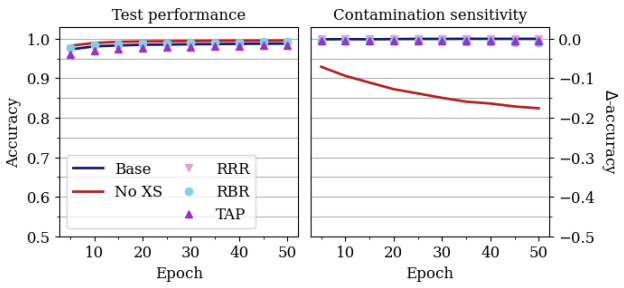

Figure 5: Test performance (accuracy) and contamination sensitivity (Δ-accuracy) for MNIST with ground-truth annotations. Higher is better.

PNEU. In Figure 6, the left panel initially suggest that the No XS model is best. The right panel, however, reveals that this is largely due to spurious signals, as the model's performance plummets when these change. The success of RRR and RBR vary with the model: RBR results in ostensibly strong performance for VGG-16, which turns out to rely on spurious signals, while RRR does so for DenseNet-121.

TAP, on the other hand, closely matches the performance of the Base model and is insensitive to the change in spurious correlations for all three CNN architectures, highlighting its efficacy for deeper models, too.

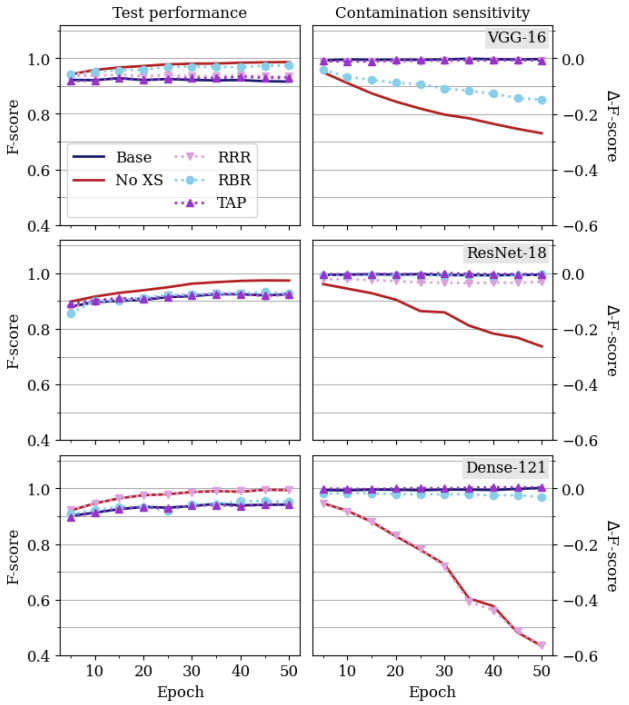

Figure 6: Test performance (F-score) and contamination sensitivity (Δ-F-score) for PNEU with ground-truth annotations. Higher is better.

KNEE. Figure 7 shows that while RRR and RBR seem to perform well (left panel) compared to TAP and even Base models, they exhibit significant contamination sensitivity (right panel) and thus only perform *spuriously* well. Similarly, the No XS model at first appears to be best-performing but in fact relies on spurious signals. TAP, in contrast, shows little contamination sensitivity and matches the Base model. These results demonstrate the ability of TAP to protect deeper CNNs from learning spurious signals.

Setting 2: Teacher Annotations

Instead of ground-truth annotations, we now assume access to a small clean dataset (10% or 1% of the contaminated data in size) on which we pre-train a teacher to obtain annotations from. We focus on TAP and leave, for completeness, results for RRR and RBR to the extended paper, which do not appear to work as well with teacher annotations compared to TAP. We also leave results on MNIST to the extended paper.

Figure 8 compares the teacher annotations to the ground-truth at different thresholds τ. The annotations are better than random (i.e., recall and precision of circa 1.9%), and raising τ increases the recall and reduces precision. This implies (i) teacher annotations can identify spurious signals albeit imperfectly, and (ii) a higher τ results in capturing more

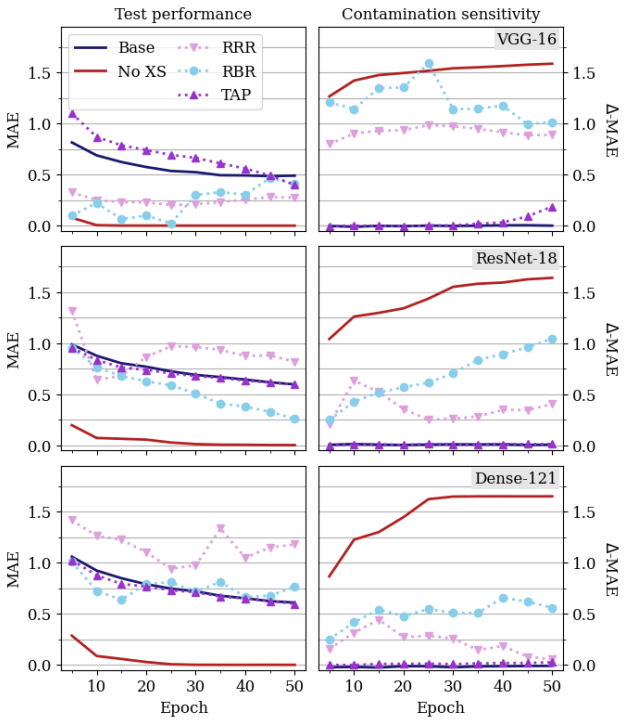

Figure 7: Test performance (MAE) and contamination sensitivity (Δ-MAE) for KNEE with ground-truth annotations. Lower is better.

spurious signals at the cost of including relevant areas. Interestingly, annotations from the 1%-teacher are almost as good as those from the 10%-teacher.

Figure 9 shows the final-epoch test performance and contamination sensitivity of using TAP with teacher and baseline random annotations. On PNEU, TAP with teacher annotations results in strong performances on the contaminated data with little sensitivity to spurious signals. Notably, there is very little difference between using the two sets of teacher annotations. The student models notably outperform both the 10%-teacher (F-score: 0.869) and 1%-teacher (F-score: 0.668). Random annotations, in contrast, result in performances which rely on spurious correlations for two of three models. This suggests that better-than-random but not necessarily perfect targeting of the spurious signals is needed.

On KNEE, the success of using teacher annotations more heavily depends on (i) τ, and (ii) the chosen model. For DenseNet-121 and VGG-16, an increase in τ surprisingly results in an increase in contamination sensitivity (a similar trend can also be observed on PNEU). This, along with the poor performance of the random annotations, suggests that noisier annotations cannot be simply offset by penalising a greater area, underscoring again the importance of good targeting. Focusing therefore on $\tau = 0.01$, we observe that only DenseNet-121 is successful in maintaining good performance which still suffers from some contamination sensitivity. VGG-16 achieves a seemingly low MAE which exploits spurious correlations. ResNet-18 did not train well with noisy annotations: while contamination sensitivity is low, this may be specious given its high MAE.

Overall, these results suggest that teacher annotations can be effective substitutes depending on the task and model. Moreover, a small clean dataset may be sufficient to pretrain a teacher that can output useful annotations.

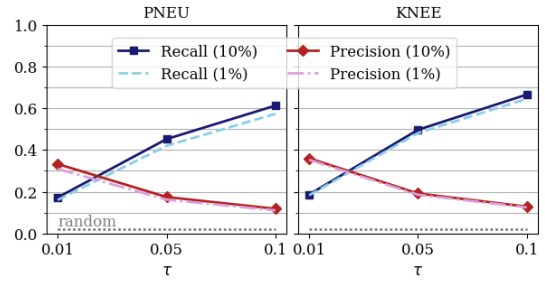

Figure 8: *Annotation* precision and recall of teachers trained on 10% or 1% clean splits at different thresholds.

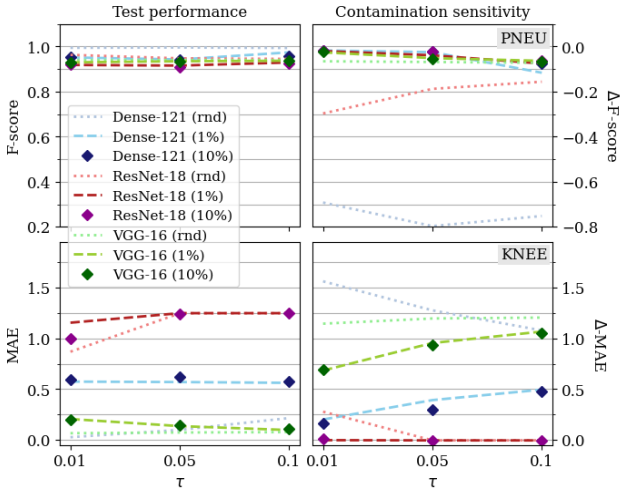

Figure 9: TAP test performance and contamination sensitivity with random (rnd) and teacher (1% or 10%) annotations vs ground-truth. Higher/Lower better for PNEU/KNEE.

Conclusion

We have proposed TAP, targeting activations of spurious signals with penalties as a way of mitigating their use in CNNs. TAP avoids expensive second-order derivatives, enabling faster training times and lower memory usage. We have shown TAP's relationship with the popular RRR and RBR losses and demonstrated its effectiveness on three datasets and four CNN architectures. We have also shown that TAP can maintain good performance with noisy teacher annotations which can help reduce otherwise costly ground-truth annotation requirements. We plan to investigate in future work if a variant of TAP can be generalised to other model architectures besides CNNs.

Acknowledgements

We thank Hamed Ayoobi, Alice Giroul and anonymous reviewers for providing feedback. Dekai Zhang was funded by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. EP/S023283/1). Matthew Williams was funded through ICHT and the Imperial/NIHR BRC. Francesca Toni was partially funded by the ERC under the EU's Horizon 2020 research and innovation programme (Grant Agreement No. 101020934) and by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme.

References

Adebayo, J.; Muelly, M.; Abelson, H.; and Kim, B. 2022. Post Hoc Explanations May Be Ineffective For Detecting Unknown Spurious Correlation. In *The Tenth International Conference on Learning Representations*.

Adler, P.; Falk, C.; Friedler, S. A.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2018. Auditing Black-box Models for Indirect Influence. *Knowledge and Information Systems*, 54(1): 95–122.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. arxiv:1910.10045.

Chen, P.; Gao, L.; Shi, X.; Allen, K.; and Yang, L. 2019. Fully Automatic Knee Osteoarthritis Severity Grading Using Deep Neural Networks with a Novel Ordinal Loss. *Computerized Medical Imaging and Graphics*, 75: 84–92.

Friedrich, F.; Stammer, W.; Schramowski, P.; and Kersting, K. 2023. A Typology for Exploring the Mitigation of Shortcut Behaviour. *Nature Machine Intelligence*, 5(3): 319–330.

Hagos, M. T.; Curran, K. M.; and Mac Namee, B. 2022. Impact of Feedback Type on Explanatory Interactive Learning. In *Foundations of Intelligent Systems*, volume 13515 of *Lecture Notes in Computer Science*, 127–137. Springer International Publishing.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. IEEE.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018*, volume 11207 of *Lecture Notes in Computer Science*, 793–811. Springer International Publishing.

Heo, J.; Piratla, V.; Wicker, M.; and Weller, A. 2023. Use Perturbations when Learning from Explanations. arXiv:2303.06419.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arxiv:1503.02531.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. IEEE.

Ke, A.; Ellsworth, W.; Banerjee, O.; Ng, A. Y.; and Rajpurkar, P. 2021. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, 116–124. Association for Computing Machinery.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; Dong, J.; Prasadha, M. K.; Pei, J.; Ting, M. Y.; Zhu, J.; Li, C.; Hewett, S.; Dong, J.; Ziyar, I.; Shi, A.; Zhang, R.; Zheng, L.; Hou, R.; Shi, W.; Fu, X.; Duan, Y.; Huu, V. A.; Wen, C.; Zhang, E. D.; Zhang, C. L.; Li, O.; Wang, X.; Singer, M. A.; Sun, X.; Xu, J.; Tafreshi, A.; Lewis, M. A.; Xia, H.; and Zhang, K. 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5): 1122–1131.e9.

Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; and Müller, K.-R. 2019. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10(1): 1096.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep Learning. *Nature*, 521(7553): 436–444.

LeCun, Y.; Cortes, C.; and Burges, C. J. 1998. The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/. Accessed: 2023-08-10.

Rajpurkar, P.; Irvin, J.; Ball, R. L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C. P.; Patel, B. N.; Yeom, K. W.; Shpanskaya, K.; Blankenberg, F. G.; Seekins, J.; Amrhein, T. J.; Mong, D. A.; Halabi, S. S.; Zucker, E. J.; Ng, A. Y.; and Lungren, M. P. 2018. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLOS Medicine*, 15(11): e1002686.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, 1135–1144. ACM.

Rieger, L.; Singh, C.; Murdoch, W. J.; and Yu, B. 2020. Interpretations Are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. In *37th International Conference on Machine Learning, ICML 2020*, 8086–8096.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations. In *IJCAI International Joint Conference on Artificial Intelligence*, 2662–2670.

Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; Shao, X.; Luigs, H. G.; Mahlein, A. K.; and Kersting, K. 2020. Making Deep Neural Networks Right for the Right Scientific Reasons by Interacting with Their Explanations. *Nature Machine Intelligence*, 2(8): 476–486.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2): 336–359.

Shao, X.; Skryagin, A.; Stammer, W.; Schramowski, P.; and Kersting, K. 2021. Right for Better Reasons: Training Differentiable Models by Constraining Their Influence Functions. In *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, volume 11A, 9533–9540. ISBN 9781713835974.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arxiv:1312.6034.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arxiv:1409.1556.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. arxiv:1412.6806.

Srinivas, S.; and Fleuret, F. 2018. Knowledge Transfer with Jacobian Matching. In *Proceedings of the 35th International Conference on Machine Learning*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Teso, S.; and Kersting, K. 2019. Explanatory Interactive Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 239–245.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–13.

Zech, J. R.; Badgeley, M. A.; Liu, M.; Costa, A. B.; Titano, J. J.; and Oermann, E. K. 2018. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLOS Medicine*, 15(11): e1002683.