From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space

Maximilian Dreyer[*1], **Frederik Pahde**[*1], **Christopher J. Anders**[2], **Wojciech Samek**[1,2,3], **Sebastian Lapuschkin**[1]

[1] Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[2] Technical University of Berlin, 10587 Berlin, Germany
[3] BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
{wojciech.samek,sebastian.lapuschkin}@hhi.fraunhofer.de

Abstract

Deep Neural Networks are prone to learning spurious correlations embedded in the training data, leading to potentially biased predictions. This poses risks when deploying these models for high-stake decision-making, such as in medical applications. Current methods for post-hoc model correction either require input-level annotations which are only possible for spatially localized biases, or augment the latent feature space, thereby *hoping* to enforce the right reasons. We present a novel method for model correction on the concept level that *explicitly* reduces model sensitivity towards biases via gradient penalization. When modeling biases via Concept Activation Vectors, we highlight the importance of choosing robust directions, as traditional regression-based approaches such as Support Vector Machines tend to result in diverging directions. We effectively mitigate biases in controlled and real-world settings on the ISIC, Bone Age, ImageNet and CelebA datasets using VGG, ResNet and EfficientNet architectures. Code and Appendix are available on https://github.com/frederikpahde/rrclarc.

1 Introduction

For over a decade, Deep Neural Networks (DNNs) face a growing interest in industry and research, featuring application in fields such as medicine or autonomous driving due to their strong predictive performance. However, their high performance may potentially be inflated by *spurious correlations* in the training data, which can pose serious risks in safety-critical applications (Geirhos et al. 2020). Several so called "short-cuts" have been found in medical settings, including hospital tags in COVID-19 radiographs (DeGrave, Janizek, and Lee 2021), or skin markings for skin lesion detection (Cassidy et al. 2022). Such short-cuts might also compromise fairness, as shown in Figure 1, where a DNN learned to use apparel features, *i.e.*, a collar, to infer that the hair is not blonde, due to overly-present dark-haired men wearing a suit in the training dataset.

In order to reveal such spurious behavior, the field of eXplainable Artificial Intelligence (XAI) has proposed several techniques identifying irregularities in a model's global behavior (Lapuschkin et al. 2019; Bykov et al. 2023; Pahde

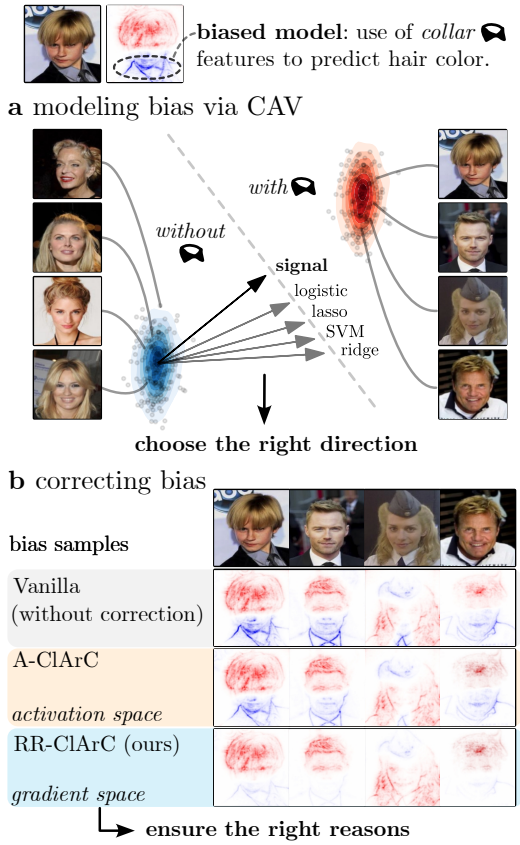

Figure 1: Our DNN bias correction method RR-ClArC consists of two steps: first finding the bias direction in the model's latent space, and secondly reducing model sensitivity towards the direction in a fine-tuning step. (*a*): When modeling the bias (here "collar") via CAVs, robust approaches such as signal-CAV are key to model correction. Most traditional regression solvers (*e.g.*, SVMs) lead to diverging directions. (*b*): Compared to activation-based methods, such as A-ClArC, that *implicitly* regularize model behavior, RR-ClArC *explicitly* reduces bias sensitivity via gradient penalization, ultimately reducing the relevance of the bias most strongly, as shown in the explanation heatmaps.

[*]These authors contributed equally.

et al. 2023), or by studying individual predictions (Teso and Kersting 2019). Acting on such findings, a variety of works perform post-hoc model correction by penalizing model attention on spurious features using pixel-wise input-level annotations (Ross, Hughes, and Doshi-Velez 2017; Rieger et al. 2020). Whereas such annotations are highly labor-intensive and only applicable for spatially localized biases, *i.e.*, biases that can clearly be located in input space, the Class Artifact Compensation (ClArC)-framework proposes bias unlearning in a model's latent feature space, requiring only sparse (sample-wise) annotations in the form of artifact labels (Anders et al. 2022). The approach follows a common methodology in the field of latent concept interpretability: Biases are modeled using latent vectors, referred to as Concept Activation Vector (CAV) (Kim et al. 2018), which, importantly, can also describe spatially unlocalized biases, such as, *e.g.*, color shifts or static artifacts from imaging equipment, that (possibly) overlay sensible input features.

We find, however, that the effectiveness of the ClArC-framework's model correction is limited by targeting only latent *activations*, which has two major drawbacks: (1) Manipulation of latent activations cannot be applied in a class-specific manner, and (2) biases may be only partially unlearned due to the method's indirect regularization.

To *enforce* the use of the right reasons on the concept level, we present Right Reason ClArC (RR-ClArC), an extension to ClArC which explicitly penalizes the model's latent gradient along the bias direction. Thus, our method enables class-specific unlearning of localized, as well as unlocalized biases while only requiring sparse sample-wise label annotation for the computation of bias CAVs. Furthermore, these annotations can be acquired semi-automatically using available XAI tools, as illustrated in (Pahde et al. 2023).

As suggested by (Kim et al. 2018), and followed in the ClArC framework, CAVs are usually computed using regression-based approaches such as Support Vector Machines (SVMs). However, we observe that common regression-based approaches hinder precise model correction caused by a tendency to diverge from the true concept direction due to, *e.g.*, noise in the data (Haufe et al. 2014). Only for the correlation-based signal-CAV (Pahde et al. 2022) we observe a strong correlation of the modeled direction and the true concept direction.

Our contributions include the following:

1. We compare different CAV computation methods and observe significant shortcomings in widely used approaches such as SVMs, which deviate strongly from the true concept direction in controlled settings.
2. We present RR-ClArC, a novel method to correct model behavior using CAVs, which is based on the latent gradient w.r.t. the model output, allowing for a class-specific unlearning of localized and unlocalized biases.
3. We evaluate the performance of RR-ClArC against other state-of-the-art methods in controlled settings on the ISIC, Bone Age and ImageNet datasets, as well as for a dataset-intrinsic bias in the CelebA dataset, using the VGG, ResNet and EfficientNet architectures.

2 Related Work

Earlier works describe the tendency of DNNs to discover shortcuts in training data (Geirhos et al. 2020), harming their generalization capabilities in real-world scenarios. Among other techniques (Robinson et al. 2021; Makar et al. 2022), XAI-based methods have proven as useful tools for the detection and removal of shortcuts learned by DNNs (Lapuschkin et al. 2019; Pahde et al. 2023; Wu et al. 2023).

Most approaches for post-hoc model correction are based on input-level guidance. These either require spatial annotations of the bias in the form of segmentation masks, which are expensive to retrieve and only applicable for localized biases (Rieger et al. 2020), or require a data augmentation of the bias to change the data distribution (Schramowski et al. 2020; Li et al. 2023). The former group of methods aims at aligning a model's attention with a pre-defined input-level prior by penalizing the use of undesired features. The popular method of Right for the Right Reason (RRR) (Ross, Hughes, and Doshi-Velez 2017), *e.g.*, achieves model correction hereby through regularization of the input gradient, introducing an additional loss term during fine-tuning.

Other recent works propose model correction on the concept-level, allowing to address biases that are not clearly localized in input space. While some are only applicable for interpretable architectures (Bontempelli et al. 2022; Yan et al. 2023), the ClArC-framework leverages CAVs to model the direction of undesired data artifacts in latent activation space (Anders et al. 2022). ClArC-methods, however, are based on latent feature augmentation and thus do not support class-specific corrections. Our approach extends the ClArC framework by *explicitly* penalizing the model for the use of artifactual data, as modeled in latent space, for the prediction of a given class, allowing class-specific corrections.

Note, that shortcut removal by data cleaning, input-level augmentation, or resampling (Zhang et al. 2018; Plumb, Ribeiro, and Talwalkar 2022; Wu et al. 2023; Li et al. 2023) is often insufficient in practice, requiring *full* re-training, where the cleaning process itself either leads to reduced training size, or is impracticably labor-intensive. In this work, we focus on post-hoc model correction based on only few fine-tuning steps.

3 Methods

We present RR-ClArC, a novel method for post-hoc model correction through the latent gradient. As illustrated in Figure 1, our method is based on two steps: (1) computing a robust CAV to model a bias concept, as described in Section 3.1, and (2) bias unlearning by penalizing model sensitivity along the CAV direction, described in Section 3.3.

A feed-forward DNN can be seen as a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, mapping input samples $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$, that is given as a composition of n functions f_i for each layer as

$$f = \underbrace{f_n \circ f_{n-1} \circ \cdots \circ f_{l+1}}_{\tilde{f}:\mathbb{R}^m \rightarrow \mathcal{Y}} \underbrace{\circ f_l \circ \cdots \circ f_1}_{\mathbf{a}:\mathcal{X} \rightarrow \mathbb{R}^m} . \quad (1)$$

We can further split the model f into two parts as noted in Equation (1): a feature extractor $\mathbf{a} : \mathcal{X} \rightarrow \mathbb{R}^m$ for the latent activations of a chosen layer l (with m neurons), and a model head $\tilde{f} : \mathbb{R}^m \rightarrow \mathcal{Y}$, mapping the activations to the outputs.

3.1 Choosing The Right Direction

The authors of (Kim et al. 2018) define a Concept Activation Vector as the normal to a hyperplane separating samples without a concept and samples with a concept in the model's latent activations for a selected layer. This hyperplane is commonly computed by solving a classification problem, *e.g.*, using SVMs, ridge, lasso, or logistic regression (Kim et al. 2018; Pfau et al. 2021; Anders et al. 2022; Yuksekgonul, Wang, and Zou 2023). We refer to Appendix A.4 for details on optimizer differences. The weight vector resulting from classification solvers, however, are not necessarily ideal CAVs, as the direction best separating two distributions does not always point from one distribution to the other (Haufe et al. 2014). To that end, signal-pattern-based CAVs ("signal-CAVs") have been proposed (Pahde et al. 2022), which are more robust against noise.

Concretely, a signal-CAV is given by the correlation between latent activations $\mathbf{a}(\mathbf{x})$ of samples $\mathbf{x}$ and concept labels $t \in \{0, 1\}$ of the concept-labeled dataset $\mathbf{x}, t \in \mathcal{X}_\mathbf{h}$ as

$$\mathbf{h}^{\text{signal}} = \sum_{\mathbf{x},t \in \mathcal{X}_\mathbf{h}} (\mathbf{a}(\mathbf{x}) - \bar{\mathbf{a}})(t - \bar{t}) \tag{2}$$

with mean activation $\bar{\mathbf{a}} = \frac{1}{|\mathcal{X}_\mathbf{h}|} \sum_{\mathbf{x},t \in \mathcal{X}_\mathbf{h}} \mathbf{a}(\mathbf{x})$ and mean concept label $\bar{t} = \frac{1}{|\mathcal{X}_\mathbf{h}|} \sum_{\mathbf{x},t \in \mathcal{X}_\mathbf{h}} t$.

3.2 Class Artifact Compensation

The ClArC framework corrects model (mis-)behavior w.r.t. an artifact by modeling its direction $\mathbf{h}$ in latent space using CAVs. The framework consists of two methods, namely Augmentive ClArC (A-ClArC) and Projective ClArC (P-ClArC). While A-ClArC adds the artifact to activations $\mathbf{a}(\mathbf{x})$ of layer l for all samples in a fine-tuning phase, hence teaching the model to become more invariant towards that direction, P-ClArC suppresses the artifact direction during the test phase and does not require fine-tuning. More precisely, the perturbed activations $\mathbf{a}'(\mathbf{x})$ are given by

$$\mathbf{a}'(\mathbf{x}) = \mathbf{a}(\mathbf{x}) + \gamma(\mathbf{x})\mathbf{h} \tag{3}$$

with perturbation strength $\gamma(\mathbf{x})$ depending on input $\mathbf{x}$. Here, $\gamma(\mathbf{x})$ is chosen such that the activation in direction of the CAV is as high as the average value over non-artifactual or artifactual samples for P-ClArC or A-ClArC, respectively.

3.3 RR-ClArC: Penalizing the Wrong Reasons

To reduce a DNN's bias sensitivity, we introduce RR-ClArC, which *explicitly* penalizes the output gradient in the direction of a bias CAV $\mathbf{h}$. Specifically, RR-ClArC introduces an additional loss term for a fine-tuning step which penalizes the model for using features aligning with the bias direction by computing the inner product between CAV $\mathbf{h}$ (*bias direction*) and the gradient of the output w.r.t. the latent features $\nabla_\mathbf{a} \tilde{f}(\mathbf{a}(\mathbf{x}))$ (*sensitivity w.r.t. latent features*), given by

$$L_{\text{RR}}(\mathbf{x}) = \left(\nabla_\mathbf{a} \tilde{f}(\mathbf{a}(\mathbf{x})) \cdot \mathbf{h} \right)^2 . \tag{4}$$

To enable a class-specific bias unlearning for models with k (multiple) output classes, described by function $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^k$, we extend the loss term to

$$L_{\text{RR}}(\mathbf{x}) = \left(\nabla_\mathbf{a} \left[\mathbf{m} \cdot \tilde{\mathbf{f}}(\mathbf{a}(\mathbf{x})) \right] \cdot \mathbf{h} \right)^2 \tag{5}$$

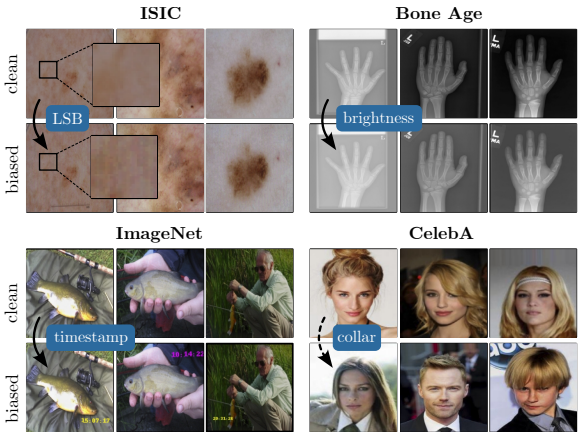

Figure 2: Overview of investigated data biases. (*Top left*): Samples of the ISIC dataset are corrupted using an LSB attack. (*Top right*): The brightness of samples from the Bone Age dataset is increased. (*Bottom left*): An artificial timestamp is added to samples from ImageNet. (*Bottom right*): The CelebA dataset intrinsically has a negative correlation between the existence of collars and blonde hair.

where the regularization can be controlled class-specifically with annotation vector $\mathbf{m} \in \mathbb{R}^k$. For regularizing all classes, we can set all elements of $\mathbf{m}$ to one. However, choosing elements uniformly randomly as $(\mathbf{m})_i \in_R \{-1, 1\}$ for each sample $\mathbf{x}$ improves regularization, further motivated in Appendix E. Alternatively, we can also correct for a specific class c by choosing $(\mathbf{m})_i = \delta_{ic}$ with Kronecker-delta δ, which ensures that related (harmless) concepts relevant for other classes are not unlearned (see Section 4.4).

Intuitively, the loss L_{RR} penalizes the model for changing the output when slightly adding or removing activations along the bias direction $\mathbf{h}$, *i.e.*

$$\lim_{\epsilon \to 0} \frac{\tilde{f}(\mathbf{a}(\mathbf{x}) + \epsilon \mathbf{h}) - \tilde{f}(\mathbf{a}(\mathbf{x}))}{\epsilon} \approx \nabla_\mathbf{a} \tilde{f}(\mathbf{a}(\mathbf{x})) \cdot \mathbf{h} \overset{!}{=} 0 . \tag{6}$$

Thus, by minimizing L_{RR}, the model becomes insensitive towards the bias direction. Note, that in order to ensure that the bias direction $\mathbf{h}$ in layer l stays constant, all weights of layers $l' \leq l$ need to be frozen during the fine-tuning phase.

4 Experiments

Our experiments aim to answer the following questions:

- **(Q1)** How well do CAVs computed with typical approaches align with the true bias direction?
- **(Q2)** How effective does RR-ClArC revise biases compared to other state-of-the-art methods?
- **(Q3)** Is RR-ClArC suitable for class-specific unlearning?
- **(Q4)** How does each component of RR-ClArC affect its performance?

4.1 Experimental Setting

Datasets and Models We fine-tune pre-trained VGG-16 (Simonyan and Zisserman 2015), ResNet-18 (He

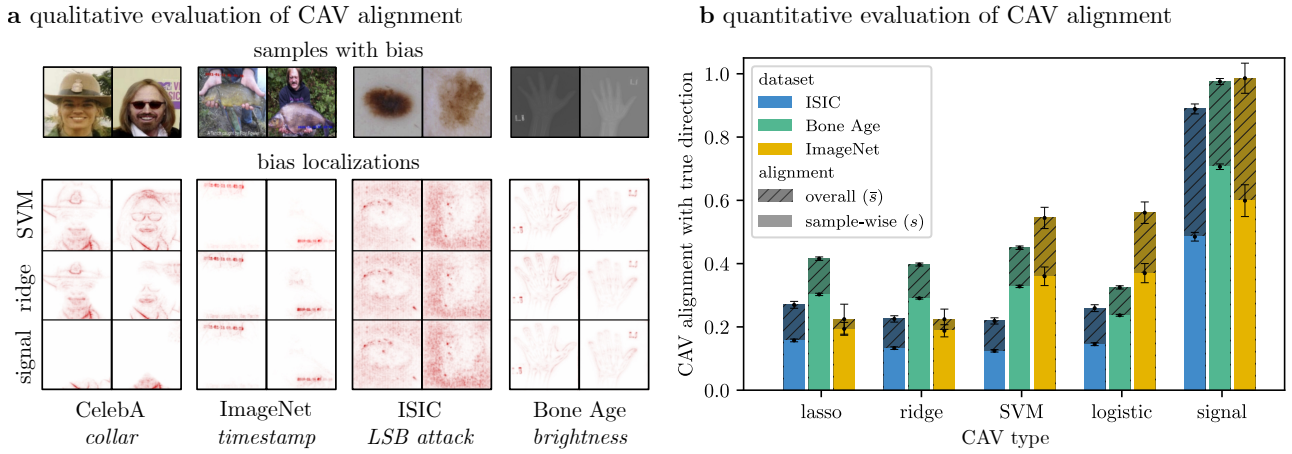

Figure 3: Evaluating CAV alignments. (*a*): CAV localization heatmaps allow to qualitatively check the alignment of a CAV with a concept. Whereas the signal-CAV tends to pinpoint the bias best for CelebA, all CAV types similarly highlight the biases of the ImageNet, ISIC, Bone Age dataset. (*b*): In controlled settings, we can compute the *true* alignment with the modeled CAV by measuring the change in activations when the bias concept is added to the input. Here, signal-CAV leads to significantly better alignments than other commonly used regression-based approaches, such as SVM-CAV in *all* experiments. This indicates, that sensible CAV localizations do not necessitate a high alignment. The Standard Error of the Mean (SEM) is shown by error bars.

et al. 2016) and EfficientNet-B0 (Tan and Le 2019) models on the ISIC 2019 dataset (Codella et al. 2018; Tschandl, Rosendahl, and Kittler 2018; Combalia et al. 2019) for skin lesion classification, the Pediatric Bone Age dataset (Halabi et al. 2019) for bone age estimation based on hand radiographs, ImageNet (Deng et al. 2009) for large scale visual recognition, and the CelebA dataset (Liu et al. 2015), offering face attributes of celebrities with the task to predict hair color. For the former three datasets, we artificially insert "Clever Hans" artifacts, *i.e.*, features unrelated to the given task, yet correlating with the target class, into data samples from only one class in a controlled setting, encouraging the model to learn a shortcut. Specifically, we use two spatially unlocalized artifacts, by (1) increasing the image brightness, i.e., increasing pixel values for each color channel, for the Bone Age dataset, and (2), run a least significant bit (LSB) attack on ISIC 2019. Inspired by steganography techniques, LSB attacks hide secret messages, *e.g.*, a text message converted into a bit stream, into the least significant bits of input features (here: voxel values) of the DNN (Li et al. 2020), which is hardly visible to the human eye. For ImageNet, we insert a timestamp to images, mimicking the option of digital cameras to add time information via text overlay. Lastly, for CelebA, we leverage a dataset-intrinsic bias, namely the negative correlation of the existence of collars and blonde hair, which is picked up by models to predict hair color (see Figure 1). An overview of the data artifacts is provided in Figure 2. All datasets are split into training, validation, and test sets. Additional dataset and training details are provided in Appendix A.

Concept Activation Vectors To compute CAVs, the last convolutional layer is chosen for all models to extract features $\mathbf{a}(\mathbf{x})$, most likely representing disentangled representations (Zeiler and Fergus 2014). We tune hyperparameters γ as $\gamma \in \{10^{-5}, 10^{-4}, \ldots, 10^{5}\}$ for all regression-based optimizers. Further details are given in Appendix A.

4.2 CAV Alignment With True Direction (Q1)

A CAV, modeling an artifact, ideally represents the direction in latent activation space pointing from a cluster of clean samples to a cluster of artifactual samples. However, the alignment between CAV and (usually unknown) concept direction is often not evaluated. In the following, we illustrate a qualitative and quantitative approach to assess CAVs, and identify shortcomings of regression-based CAV optimizers.

Qualitative Alignment To bridge the gap between latent activation space and human-interpretable input space, we follow (Anders et al. 2022; Pahde et al. 2023), and localize our biases in the input image based on CAVs. Such localization can be achieved by computing an attribution heatmap, not w.r.t. to the output logit as for standard explanation heatmaps, but w.r.t. the dot-product between CAV and activations, *i.e.*, $\mathbf{a} \cdot \mathbf{h}$. All computed heatmaps are based on the LRP attribution method (Bach et al. 2015) with the $\varepsilon z^+\flat$-composite (Kohlbrenner et al. 2020), as implemented in the `zennit` (Anders et al. 2021) package.

Examples of bias localization heatmaps using signal, ridge and SVM-CAVs are shown in Figure 3a for all experimental datasets with VGG-16. For the CelebA dataset, signal-CAVs tend to pinpoint the collar bias best, with ridge and SVM also highlighting unrelated features. Regarding ImageNet and the unlocalized biases (ISIC, Bone Age), all CAV types highlight bias features. Here, no difference between CAV types is apparent. More bias localizations are provided in Appendix B. Note, however, that a good agreement in localization does not necessitate good alignment, as the quantitative evaluation in the following section shows.

Figure 4: Effect of model correction on explanation heatmaps for the VGG-16 model on all datasets. Whereas RRR successfully decreases the relevance on localized biases (ImageNet and CelebA), RRR tends to fail on unlocalized artifacts (ISIC, Bone Age), with the model's attention focusing on arbitrary features. Overall, RR-ClArC reduces bias attention most reliably.

Quantitative Alignment In order to measure to which degree a bias CAV corresponds to the true direction in the latent space, we perform experiments in a controlled setting (ISIC, ImageNet, Bone Age). Specifically, we apply an input transformation φ that adds the bias to the input as $\varphi(\mathbf{x}) = \tilde{\mathbf{x}}$, and measure the cosine similarity between the bias CAV and the resulting difference in activations $\Delta\mathbf{a}(\mathbf{x}) = \mathbf{a}(\tilde{\mathbf{x}}) - \mathbf{a}(\mathbf{x})$. The alignment s of CAV $\mathbf{h}$ is then given as

$$s = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\mathbf{h} \cdot \Delta\mathbf{a}(\mathbf{x})}{\|\mathbf{h}\|_2 \|\Delta\mathbf{a}(\mathbf{x})\|_2}, \tag{7}$$

where the alignment is computed w.r.t. to the activation difference in *each* sample of dataset $\mathcal{X}$. Alternatively, we also compute the overall alignment $\bar{s}$ given by

$$\bar{s} = \frac{\mathbf{h}}{\|\mathbf{h}\|_2} \cdot \frac{\frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \Delta\mathbf{a}(\mathbf{x})}{\| \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x}' \in \mathcal{X}} \Delta\mathbf{a}(\mathbf{x}') \|_2}, \tag{8}$$

which describes how well the CAV aligns with the mean activation change over *all* samples.

The resulting alignment scores are shown in Figure 3b for the VGG-16 model and the ISIC, ImageNet and Bone Age datasets. Here, we compare CAVs computed through SVM, ridge, lasso and logistic regression as well as correlation (signal). It is apparent, that signal-CAVs perform significantly better than the other approaches, resulting in overall alignments $\bar{s}$ of over 90 % compared to less than 60 % for the regression CAVs. Also, for the sample-wise alignment s, signal-CAVs often shows twice as high alignment scores as other approaches. For EfficientNet and ResNet, similar results can be observed, as shown in Appendix B.

The sample-wise alignment s is generally smaller than the overall alignment. We find that transformation φ barely affects samples in some cases, *e.g.*, when adding a black timestamp to a dark background (ImageNet), leading to low alignments s. We show samples with high and low alignment for all experiments in Appendix Figure A.3.

All in all, the qualitative and quantitative CAV evaluations show, that the CAV optimizer can have a significant impact on the alignment, with regression-based optimizers showing diverging directions. Sensible bias localizations may hint at diverging directions, but do not necessitate good alignment, underlined by the quantitative evaluation.

4.3 Revising Model Biases (Q2)

We revise model biases by applying the methods of A- and P-ClArC, the input-level correction method RRR, as well as our introduced method RR-ClArC, and compare with a *Vanilla* model. All models are fine-tuned for 10 epochs. Although not requiring further training, before the application of P-ClArC we fine-tune the model in Vanilla fashion for better comparability. For RRR and RR-ClArC, we test different regularization strengths λ, detailed in Appendix A.3. As RRR requires prior input-level bias localizations, we use a threshold to convert artifact localizations retrieved as described in Section 4.2 into binary masks for CelebA. In the controlled experiments, we generate ground truth binary masks to localize inserted artifacts. For unlocalized artifacts (spreading over the entire input) ground truth masks cover the full image. We use signal-CAVs to represent artifacts in latent space for ClArC-methods, as they have shown the best alignment scores in Section 4.2. We fine-tune models using the training set, choose optimal λ values using the validation set, and measure the accuracy on a *clean* and a *biased* test set. While we insert the Clever Hans artifact into all samples from the *biased* test set in the controlled setting, for CelebA we use a subset from the original test set with samples containing the collar-artifact.

Moreover, we measure the remaining sensitivity on the

architecture	method	Bone Age			ISIC			ImageNet			CelebA		
		clean	*biased*	TCAV	*clean*	*biased*	TCAV	*clean*	*biased*	TCAV	*clean*	*biased*	TCAV
VGG-16	*Vanilla*	78.8	49.8	0.86	76.2	34.9	0.84	68.7	43.5	0.63	93.7	82.8	0.37
	RRR	78.8	49.8	0.86	76.7	42.8	0.72	68.6	49.6	0.55	93.7	91.2	0.43
	P-ClArC	78.9	77.4	0.66	75.1	49.0	0.77	68.3	**62.6**	0.37	56.6	60.8	0.19
	A-ClArC	77.8	69.0	0.66	75.2	49.5	0.65	67.7	60.9	**0.49**	93.0	90.4	0.44
	RR-ClArC (ours)	78.8	**77.7**	**0.52**	74.3	**57.0**	**0.49**	68.5	**62.6**	**0.49**	93.6	**92.6**	**0.54**
ResNet-18	*Vanilla*	75.1	46.3	1.00	81.8	56.8	1.00	66.7	52.9	1.00	96.8	58.3	0.21
	RRR	74.5	47.9	1.00	78.7	61.1	1.00	66.4	59.1	0.08	95.5	74.7	0.92
	P-ClArC	75.0	70.7	**0.60**	60.8	59.9	1.00	67.0	61.7	0.80	96.5	64.4	0.06
	A-ClArC	74.8	57.4	0.34	77.1	65.0	0.98	65.0	63.3	0.88	96.1	62.9	0.38
	RR-ClArC (ours)	71.1	**74.2**	0.39	78.5	**71.2**	**0.76**	66.5	**64.0**	**0.55**	95.8	**75.3**	**0.61**
Efficient Net-B0	*Vanilla*	78.2	44.3	0.90	84.2	62.9	1.00	73.9	53.2	0.99	96.6	58.3	0.25
	RRR	78.4	49.6	0.79	83.1	68.7	0.85	73.9	59.1	0.66	95.4	75.6	**0.50**
	P-ClArC	65.2	35.1	0.02	19.7	29.6	1.00	74.1	54.6	0.21	96.8	55.0	0.05
	A-ClArC	78.0	54.2	0.64	77.7	72.8	0.68	71.4	69.9	0.90	96.7	60.6	0.24
	RR-ClArC (ours)	77.6	**70.3**	**0.53**	78.7	**75.6**	**0.54**	73.9	**70.8**	**0.56**	92.0	**77.6**	0.43

Table 1: Model correction results for all experiments. We report model accuracy (in %) on *clean* and *biased* test sets, as well as TCAV bias score. Higher scores are better for accuracy and scores close to 0.5 are best for TCAV, with best scores bold.

bias in latent space via TCAV (Kim et al. 2018), given as

$$\text{TCAV} = \frac{\left|\{\mathbf{x} \in \mathcal{X}_{\text{bias}} : \boldsymbol{\nabla}_{\mathbf{a}} \tilde{f}(\mathbf{a}(\mathbf{x})) \cdot \mathbf{h} > 0\}\right|}{|\mathcal{X}_{\text{bias}}|} \tag{9}$$

computed over the set of all bias samples $\mathcal{X}_{\text{bias}}$. For a bias-insensitive model, we expect TCAV ≈ 0.5, describing a random bias impact. For a positively or negatively contributing bias, we expect TCAV > 0.5 or TCAV < 0.5, respectively.

Results The results for all architectures and datasets are shown in Table 1, with respective standard errors given in Appendix C. Across all experiments, RR-ClArC outperforms all competitors on the *biased* test set in terms of accuracy, while not significantly hurting the classification performance on the *clean* test set. Here, RR-ClArC shows the best tradeoff between accuracy on clean and biased data. Note, that an accuracy drop on the clean test set can be expected, as the bias concept might be (partially) entangled with sensible related concepts in latent space, which, consequently, are suppressed during unlearning as well. However, in principle, the clean accuracy could also increase when alternative strategies based on other features are found.

Notably, RRR increases the biased test set accuracy only for localized biases effectively (ImageNet, CelebA). This is expected, as input localizations for unlocalized biases cover the full image (including sensible features), and thus tend to steer the models' attention towards sparse, but (possibly) insensible features. Our findings are supported by qualitative results in Figure 4, showing LRP explanation heatmaps for the corrected and Vanilla models. While the Vanilla models show large fractions of relevance on the biases (background for unlocalized artifacts), RR-ClArC performs best in teaching the model to solely focus on the object of interest, which RRR can only achieve for the localized artifacts.

Whereas ClArC-methods operating in latent activation space show reasonable accuracy gains in comparison with the Vanilla model for most tasks, they mostly only slightly decrease bias sensitivity. RR-ClArC yields better results for TCAV, *i.e.* TCAV ≈ 0.5, due to the explicit penalization of bias sensitivity in latent space, as described in Equation (5). Note, that TCAV only takes into account the sign of bias sensitivity, not the magnitude. To that end, we further report bias sensitivity magnitudes in Appendix C, and alternatively, also report input-level bias relevances for localized artifacts.

Overall, RR-ClArC yields the most reliable results, recovering the classification performance on biased datasets for both localized and unlocalized artifacts, while strongly decreasing bias sensitivity and remaining predictive performance on the clean test set.

Computational Cost By performing a single forward pass per sample, A-ClArC is as computationally expensive as Vanilla training. RR-ClArC increases training time for the VGG-16 model by about 20 % due to, *e.g.*, additionally requiring a partial backward pass (up to the CAV layer) to compute the latent gradient for the loss. RRR is most expensive with a time increase of about 73 % for VGG-16, requiring both a full forward and backward pass to compute the loss. Note, that P-ClArC does not require fine-tuning. Exact training times are given in Table A.6 of the Appendix.

4.4 Class-specific Model Correction (Q3)

Another advantage of RR-ClArC in comparison to other ClArC methods is its ability to correct model behavior w.r.t. bias concepts for *specific* classes by specifying annotation vector $\mathbf{m}$ in Equation 5 accordingly, penalizing the gradient only for chosen classes. This allows to teach the model to ignore concepts, e.g., the timestamp artifact (ImageNet), for a specific class (here: "tench"), while allowing other classes such as, *e.g.*, "digital clock", to rely on related concepts.

In this experiment, we study the impact of model correc-

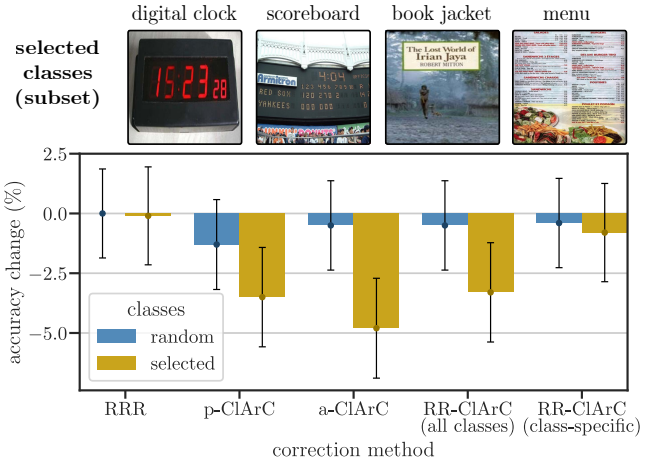

Figure 5: Impact of model correction for the timestamp bias (ImageNet) on model accuracy for random and selected classes (*bottom*), where selected classes are highly impacted by the timestamp bias, *e.g.*, "digital clock" (*top*). While A-ClArC and P-ClArC are class-inspecific and lead to a drop in accuracy for selected classes, RRR and RR-ClArC *only* unlearn the artifact for a specific class. Note, that RR-ClArC can target both, all classes or specific classes.

tions on model accuracy w.r.t. classes using concepts related to the timestamp bias. A *selected* subset of related classes is identified by a strong increase in their output logit when adding a timestamp to the input, listed in Appendix D. We then compare the impact of model corrections on the accuracy for clean samples between the *selected* classes and *all* classes. The change in accuracies for VGG-16 corrections in comparison with the Vanilla model is shown in Figure 5, and for other architectures in Appendix D. As expected, A- and P-ClArC lead to a significant drop in accuracy for the *selected* classes, confirming that the model suppressed related concepts required to recognize these classes. RRR and RR-ClArC, however, targeting the "tench" class only through the gradient, retain the model's accuracy for the *selected* classes. Note, that RR-ClArC can also target *all* classes, showing similar results as A- and P-ClArC then.

4.5 Ablation Study (Q4)

The following ablation study measures the impact of the main influencing factors of RR-ClArC, including the choice of CAV optimizer and regularization strength. Other factors, including the number of fine-tuning epochs and the choice of the target w.r.t. which the gradient is computed as in Equation (5), are discussed in Appendix E.

Concept Activation Vector When comparing the accuracy on the biased test set after model correction with RR-ClArC using different CAV optimizers, signal-CAVs significantly outperform all competitors, as shown in Figure 6 for the VGG-16 model with controlled biases. This follows the trend in the alignment experiment of Section 4.2, confirming that a high CAV alignment is required to effectively regularize all parts of a concept to fully unlearn harmful concepts.

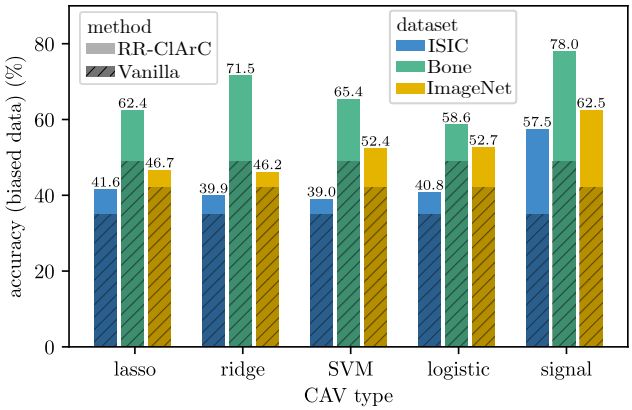

Figure 6: On all experiments, the CAV type has a significant impact on model correction (regarding accuracy on the biased dataset), with signal-CAV leading to the highest accuracies (Standard Error (SE) less than 1 %). These results mirror the CAV alignment experiment in Figure 3b.

Regularization Strength Besides CAV type, the regularization strength λ is another important parameter controlling the amount of unlearning. The higher λ, the higher the accuracy on the biased dataset up to a turning point, when the regularization becomes too strong, and the overall accuracy is reduced again, also shown in Figure A.18 of the Appendix. When a CAV is not perfectly aligned, or the bias concept entangled with useful directions in the latent space, strong regularization can be expected to harm performance.

5 Conclusion

We present RR-ClArC, a post-hoc model correction method based on CAVs and the latent gradient. RR-ClArC is a step towards easier and more generalized model correction requiring only sparse labels to unlearn any concept (unlocalized or localized) class-specifically, and being applicable to any DNN architecture with access to latent features. Throughout experiments with three popular DNN architectures on four datasets with controlled and data-intrinsic biases, RR-ClArC unlearns biases most effectively and consistently compared to other state-of-the-art approaches. In our experiments, we find that Concept Activation Vectors, which are usually applied to model latent concepts, tend to result in diverging directions when based on popular regression-based approaches such as, *e.g.*, SVMs. An important future direction will be to investigate these shortcomings further, with the possibility to improve concept-based methods in various applications.

Limitations RR-ClArC requires the freezing of layers from input to the feature layer where the bias-CAV is modeled. However, in principle, a subsequent fine-tuning step without freezing parameters on clean data is possible. Moreover, our experiments show that a well-aligned CAV is necessary for effective bias unlearning. Choosing the last convolutional layer for modeling CAVs might not always be optimal. It is still an open question, how to choose the optimal layer for bias correction.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research (BMBF) as grant BIFOLD (01IS18025A, 01IS180371I); the German Research Foundation (DFG) as research unit DeSBi (KI-FOR 5363); the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant TEMA (101093003); the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant iToBoS (965221); and the state of Berlin within the innovation support programme ProFIT (IBB) as grant BerDiBa (10174498).

References

Anders, C. J.; Neumann, D.; Samek, W.; Müller, K.-R.; and Lapuschkin, S. 2021. Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. arXiv:2106.13200.

Anders, C. J.; Weber, L.; Neumann, D.; Samek, W.; Müller, K.-R.; and Lapuschkin, S. 2022. Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models. *Information Fusion*, 77: 261–295.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one*, 10(7): e0130140.

Bontempelli, A.; Teso, S.; Giunchiglia, F.; and Passerini, A. 2022. Concept-level Debugging of Part-Prototype Networks. In *Workshop on Trustworthy Artificial Intelligence as a part of the ECML/PKDD 22 program*.

Bykov, K.; Deb, M.; Grinwald, D.; Muller, K. R.; and Höhne, M. M. 2023. DORA: Exploring Outlier Representations in Deep Neural Networks. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*.

Cassidy, B.; Kendrick, C.; Brodzicki, A.; Jaworek-Korjakowska, J.; and Yap, M. H. 2022. Analysis of the ISIC Image Datasets: Usage, Benchmarks and Recommendations. *Medical image analysis*, 75: 102305.

Codella, N. C.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *15th International Symposium on Biomedical Imaging (ISBI 2018)*, 168–172. IEEE.

Combalia, M.; Codella, N. C.; Rotemberg, V.; Helba, B.; Vilaplana, V.; Reiter, O.; Carrera, C.; Barreiro, A.; Halpern, A. C.; Puig, S.; et al. 2019. BCN20000: Dermoscopic Lesions in the Wild. arXiv:1908.02288.

DeGrave, A. J.; Janizek, J. D.; and Lee, S.-I. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7): 610–619.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Halabi, S. S.; Prevedello, L. M.; Kalpathy-Cramer, J.; Mamonov, A. B.; Bilbily, A.; Cicero, M.; Pan, I.; Pereira, L. A.; Sousa, R. T.; Abdala, N.; et al. 2019. The RSNA pediatric bone age machine learning challenge. *Radiology*, 290(2): 498–503.

Haufe, S.; Meinecke, F.; Görgen, K.; Dähne, S.; Haynes, J.-D.; Blankertz, B.; and Bießmann, F. 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87: 96–110.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*, 2668–2677. PMLR.

Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; and Lapuschkin, S. 2020. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

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.

Li, S.; Xue, M.; Zhao, B. Z. H.; Zhu, H.; and Zhang, X. 2020. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5): 2088–2105.

Li, Z.; Evtimov, I.; Gordo, A.; Hazirbas, C.; Hassner, T.; Ferrer, C. C.; Xu, C.; and Ibrahim, M. 2023. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20071–20082.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Makar, M.; Packer, B.; Moldovan, D.; Blalock, D.; Halpern, Y.; and D'Amour, A. 2022. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, 739–766. PMLR.

Pahde, F.; Dreyer, M.; Samek, W.; and Lapuschkin, S. 2023. Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models. In Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S.; Duncan, J.; Syeda-Mahmood, T. F.; and Taylor, R. H., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Proceedings, Part II*, volume 14221 of *Lecture Notes in Computer Science*, 596–606. Springer.

Pahde, F.; Weber, L.; Anders, C. J.; Samek, W.; and Lapuschkin, S. 2022. PatClArC: Using pattern concept activation vectors for noise-robust model debugging. arXiv:2202.03482.

Pfau, J.; Young, A. T.; Wei, J.; Wei, M. L.; and Keiser, M. J. 2021. Robust semantic interpretability: Revisiting concept activation vectors. arXiv:2104.02768.

Plumb, G.; Ribeiro, M. T.; and Talwalkar, A. 2022. Finding and Fixing Spurious Patterns with Explanations. *Transactions on Machine Learning Research*, 2022.

Rieger, L.; Singh, C.; Murdoch, W.; and Yu, B. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, 8116–8126. PMLR.

Robinson, J.; Sun, L.; Yu, K.; Batmanghelich, K.; Jegelka, S.; and Sra, S. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34: 4974–4986.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th 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.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y.; and LeCun, Y., eds., *International Conference on Learning Representations, ICLR 2015*.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

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.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Wu, S.; Yuksekgonul, M.; Zhang, L.; and Zou, J. 2023. Discover and Cure: Concept-aware Mitigation of Spurious Correlation. In *International Conference on Machine Learning*.

Yan, S.; Yu, Z.; Zhang, X.; Mahapatra, D.; Chandra, S. S.; Janda, M.; Soyer, P.; and Ge, Z. 2023. Towards Trustable Skin Cancer Diagnosis via Rewriting Model's Decision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11568–11577.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2023. Post-hoc Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations*.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, 818–833. Springer.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.