EXPLORING UNFAIRNESS IN INTEGRATED GRADIENTS BASED ATTRIBUTION METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Numerous methods have attempted to explain and interpret predictions made by machine learning models in terms of their inputs. Known as "attribution methods" they notably include the Integrated Gradients method and its variants. These are based upon the theory of Shapley Values, a rigorous method of fair allocation according to mathematical axioms. Integrated Gradients has axioms derived from this heritage with the implication of a similar rigorous, intuitive notion of fairness. We explore the difference between Integrated Gradients and more direct expressions of Shapley Values in deep learning and find Integrated Gradients' guarantees of fairness weaker; in certain conditions it can give wholly unrepresentative results. Integrated Gradients requires a choice of "baseline", a hyperparameter that represents the 'zero attribution' case. Research has shown that baseline choice critically affects attribution quality, and increasingly effective baselines have been developed. Using purpose-designed scenarios we identify sources of inaccuracy both from specific baselines and inherent to the method itself, sensitive to input distribution and loss landscape. Failure modes are identified for baselines including Zero, Mean, Additive Gaussian Noise, and the state of the art Expected Gradients. We develop a new method, Integrated Certainty Gradients, that we show avoids the failures in these challenging scenarios. By augmenting the input space with "certainty" information, and training with random degradation of input features, the model learns to predict with varying amounts of incomplete information, supporting a zero-information case which becomes a natural baseline. We identify the axiomatic origin of unfairness in Integrated Gradients, which has been overlooked in past research.

1 INTRODUCTION

Attribution, the identification of the input features most salient to a model prediction, is an increasingly important requirement for neural networks. It is a core part of model interpretability, which is valuable as a research tool and design aid, but also increasingly as an output requirement in its own right for uses as diverse as medical imaging (Singh et al., 2020) to loan applications (Bhatt et al., 2020). A "right to explanation" of machine decisions is even provided in the European Union's General Data Protection Regulation (Goodman & Flaxman, 2017).

Gradients of predictions with respect to model inputs can be calculated using backpropagation and have been used for feature attribution ("Vanilla Gradient") (Erhan et al., 2009; Simonyan et al., 2014; Yosinski et al., 2015). The gradients indicate which features are most sensitive to a perturbation, causing the largest change in prediction outcome. Empirically, this method often does highlight areas relevant to prediction, but it may also miss important areas due to the nonlinearity of the model and saturation of the gradients (Sundararajan et al., 2016).

The method of *Integrated Gradients* (Sundararajan et al., 2016; 2017) overcomes these issues by generating attributions based on the theory of Shapley values (Sundararajan et al., 2017). The method integrates the gradients as the input varies linearly between a *baseline* and the final input of interest:

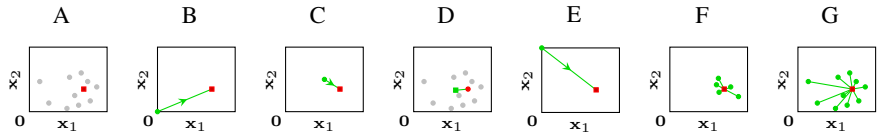


Figure 1: Examples of (generalized) baselines. Grey: input distribution; Red: attributed input; Green: baselines and integration paths. A: Distribution and target input. Baselines: B: Zero; C: Middle; D: Distribution mean; E: Maximum distance; F: Additive Gaussian noise; G: Expected Gradients. Failure modes: B-D are single static baselines so exhibit baseline blindness. Maximum Distance gives poor attributions due to extreme O.O.D. baselines. AGN fails the Local Minimum scenario (section 5.2). EG fails the Burnt Snacks scenario (section 5.1).

$$\text{Integrated Gradients}(F, \boldsymbol{x}', \boldsymbol{x}, i) := \underbrace{(\boldsymbol{x}_i - \boldsymbol{x}'_i)}_{\text{distance}} \underbrace{\int_{\alpha=0}^{1} \frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\Big|_{\boldsymbol{x} = \boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}')} d\alpha}_{\text{mean gradient}} \tag{1}$$

where $F(\cdot)$ denotes the prediction of the model, $\boldsymbol{x}$ is the input vector of interest, $\boldsymbol{x}'$ is the baseline input vector, and i indexes to the feature of interest. A numerical approximation is normally used to calculate the integral in practice. The baseline represents a *missing* or *neutral* input, a concept required by the Shapley values theory, but shared with other attribution methods (Covert et al., 2020).

In the remainder of this paper, we first give an overview of baselines discussed in the literature (Section 2). Then we discuss the theoretical cause of a failure mode of IG (*attribution transfer*) using a zero baseline example (Section 3). In Section 4 we introduce a new method, *Integrated Certainty Gradients*. Finally we show in experiments that other baselines including the state of the art Expected Gradients do not prevent attribution transfer, while in the tested scenarios Integrated Certainty Gradients does, and discuss additional failure cases of specific baselines (Section 5).

Note on notation It will be convenient to imply an equivalence between attribution methods which take a baseline, $\varphi(F, \boldsymbol{x}', \boldsymbol{x}, i)$, and those which do not $\varphi(F, \boldsymbol{x}, i)$. A baseline free method can be adapted to take a baseline by $\varphi_B(F, \boldsymbol{x}', \boldsymbol{x}, i) = \varphi(F, \boldsymbol{x}, i) - \varphi(F, \boldsymbol{x}', i)$.

2 Choice of baseline

Effective choice of baselines in practice is a topic of ongoing research. A recent review of four state-of-the-art models for tabular data did not find a best performing choice (Haug et al., 2021). As this paper noted, the concept of missingness within an arbitrary space is domain-specific. Therefore applying attribution techniques to a new domain requires this concept to be determined. Another review came to similar conclusions for image classification attribution (Sturmfels et al., 2020).

Baselines can be generalized from single inputs to distributions by taking the expected value of Integrated Gradients using each input in the distribution. In practice X' is often from an empirical (discrete uniform) distribution so the expectation becomes simply the mean over X'. A summary of common baselines used in Integrated Gradients and other attribution methods follows. For a more comprehensive review, one can refer to Haug et al. (2021); Sturmfels et al. (2020). Visualizations of integration paths are shown in Figure 1.

Constant value baselines include the uniform black/zero baseline proposed with the original method (Sundararajan et al., 2016) and choices that attempt to be unbiased with respect to the input distribution including the component-wise mean of the input distribution, the middle of the input range, and uniform noise. If the input matches the baseline, the $\boldsymbol{x}_i - \boldsymbol{x}'_i$ term in Equation 1 disappears, any any such areas will not be attributed, even when they are essential to the model's prediction ("baseline blindness") (Sturmfels et al., 2020). When the baseline is constant this cannot be avoided. This behavior can be seen clearly in attributions of MNIST images with a zero baseline, for example in Adebayo et al. (2020).

Distribution baselines avoid baseline blindness by combining multiple attributions. Uniform noise has been used in this way, although concerns have been identified about the risk of interaction between the baseline and high frequency features (Sundararajan et al., 2017; Sturmfels et al., 2020). The recently developed Expected Gradients method (Erion et al., 2020) uses the input (typically training) distribution as the baseline. It computes the attribution for a given input example efficiently by sampling baselines randomly from the training distribution, and interpolation weights from the the 0-1 range (Monte Carlo integration), rather than using the many equally spaced interpolations of the more common Riemann Sum method.

Dynamic baselines change based on the attribution input. The maximum distance baseline uses the farthest point in the input space from the input vector (Sturmfels et al., 2020). This avoids baseline blindness but abandons the concept of neutrality. Sturmfels et al. (2020) cite Fong & Vedaldi (2017) as the inspiration for blurring as a possible method of feature ablation to produce a baseline. Gaussian noise may be added to the input vector during attribution and when averaged over several iterations, the resulting attributions were found to be more detailed and less noisy (Smilkov et al., 2017). Adding noise to input vectors during training was also found to give a similar and cumulative effect. The combination of input vector with noise can be used by itself as a dynamic distribution baseline (Sturmfels et al., 2020) or as an augmentation technique in combination with another baseline. The justification of this method for use with attribution appears to be mainly empirical, and a question regarding the theoretical justification has been raised (Sturmfels et al., 2020).

3 ATTRIBUTION FAIRNESS

Assessment of attribution methods is not currently straightforward, despite ongoing research, because there is not yet a unified general procedure to do so. Rather, there are a number of tests for various indicative characteristics (Hooker et al., 2019; Adebayo et al., 2018; Yang & Kim, 2019). We focus our attention on fairness: whether the relative attributions of input components are reasonable. This concern is the central principle of Shapley Values, a celebrated method for allocating value in a game theory context, with a strongly justified mathematical foundation (Shapley, 1952), and the inspiration for Integrated Gradients methods (Sundararajan et al., 2017).

Assessing fairness is not trivial. As a reference we consider two other attribution methods defined using the original Shapley Values procedure: BShap (Sundararajan & Najmi, 2020) and SHAP (Lundberg & Lee, 2017). Shapley Values is a procedure $\phi_{\text{sv}} : (\mathcal{P}(S) \to \mathbb{R}) \to (S \to \mathbb{R})$, where $\mathcal{P}(S)$ is the power set of a set S. That is, it requires a value for each possible subset of elements, and gives back values for the elements themselves (the "Shapley values"). It does this in the unique way such that the values of all elements sum to the value of the complete set: $v(S) = \sum_{e \in S} \phi_{\text{sv}}(v, e)$, and certain axioms, widely considered as naturally related to fairness, are satisfied. As such it gives a way of sharing the value of the complete set fairly among the elements, based on the values that are achieved by each subset. To apply Shapley Values in areas not relating to sets, the inclusion or exclusion of elements may be replaced with any other binary property. In the case of BShap, the roles of sets and elements are taken by vectors and components, with 'included elements' becoming components drawn from a target vector $\boldsymbol{x}$ and 'excluded elements' becoming components drawn from a baseline vector $\boldsymbol{x}'$. Each possible assembled vector fulfills the role of a subset and the model's prediction F is used as the value function v. The Shapley values derived in this way are the attributions for the components of the target vector $\boldsymbol{x}$. SHAP uses an expectation of the model's prediction $\mathbb{E} \circ F$ as the value function v. The role of the subsets is provided by random input vectors $\mathbf{a}$ where each component $\mathbf{a}_i$ is either fixed to $\boldsymbol{x}_i$ or taken from a random input vector conditioned on the fixed components. In summary:

$\phi_{\text{sv}}(v, e)$	v	sets	elements	S	$\{\}$
$\varphi_{\text{BShap}}(F, \boldsymbol{x}', \boldsymbol{x}, i)$	F	vectors	components	$\boldsymbol{x}$	$\boldsymbol{x}'$
$\varphi_{\text{SHAP}}(F, \boldsymbol{x}, i)$	$\mathbb{E} \circ F$	random vectors	components	$\mathbf{a} = \boldsymbol{x}$	$\mathbf{a} \sim \mathbf{X}$

3.1 ATTRIBUTION TRANSFER

Integrated Gradients integrates partial derivatives along a straight path between a baseline and the input vector. The can be roughly imagined as accumulating the contours $\partial f(\boldsymbol{x})$ of the input-output landscape that fall across the integration path. To explore fairness in IG we consider the scenario

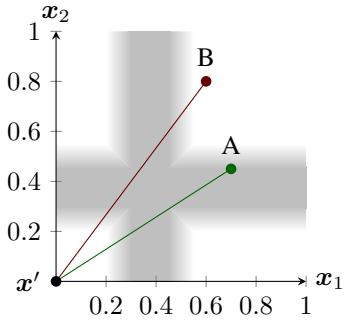


Figure 2: Representation of a model that indicates whether either component of the input $\boldsymbol{x} \in \mathbb{R}^2$ falls within the approximate range 0.25 - 0.5. The background color indicates output value, from white (false, 0) to grey (true, 1). Two IG integration paths exhibiting attribution transfer are shown. Both paths start from baseline $\boldsymbol{x}' = (0, 0)$. Path A indicates an input $\boldsymbol{x}$ with $\boldsymbol{x_2}$ in the target range and $\boldsymbol{x_1}$ outside. However since the path crosses only a 'horizontal' gradient IG will incorrectly attribute the result (class 1) to $\boldsymbol{x_1}$. Path B indicates an input $\boldsymbol{x}$ with both components above the target range. As the path crosses perpendicular gradients, both components are attributed with opposite signs.

Table 1: Attribution results associated with Figure 2. IG and BShap have baseline at (0, 0). SHAP values have a different sum because SHAP effectively uses the expectation across the whole input space as the baseline. The input distribution is assumed uniform for SHAP.

	BShap	SHAP	IG
A	(0, 1)	(-0.10, 0.66)	(1, 0)
B	(0, 0)	(-0.22, -0.22)	(-1, 1)

shown in Figure 2. A model must indicate whether either of two components $\boldsymbol{x}_1, \boldsymbol{x}_2$ fall within the range 0.25 - 0.5. Assuming a one-hot encoding of the categorical result, this is modeling a function $\mathbb{R}^2 \to \mathbb{R}^2$.

Path A shows an IG attribution path from a zero baseline (starting at the origin) to an input vector $\boldsymbol{x}$ with $\boldsymbol{x}_1$ outside the target range and $\boldsymbol{x}_2$ in the target range. Attribution results are shown in Table 1. BShap and SHAP both attribute the high output at $\boldsymbol{x}$ to $\boldsymbol{x}_2$. Because the integration path crosses only 'vertical' contours IG attributes it entirely to $\boldsymbol{x}_1$. The IG attribution is despite the change in $\boldsymbol{x}_1$ not affecting $f(\boldsymbol{x})$ at either value of $\boldsymbol{x}_2$, and the change in $\boldsymbol{x}_2$ giving a full increase in $f(\boldsymbol{x})$ for either value of $\boldsymbol{x}_1$. This result for IG could arguably be considered 'maximally misleading'.

To understand why IG gives this result, we compare the scenario with a case where there is a ground truth attribution. For additive models $f(\boldsymbol{x}) = \sum_i f_i(\boldsymbol{x}_i)$ each component's contribution can be considered independently: $\varphi(f, \boldsymbol{x}', \boldsymbol{x}, i) = f_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}'_i)$. Linear models are a subset. In this regime BShap and IG give matching results, as does SHAP if the baselines $\boldsymbol{x}'$ were chosen such that $f_i(\boldsymbol{x}'_i) = \mathbb{E}(f_i)$ and the input distribution is uniform. The model in Figure 2 deviates from additivity where the target regions of each component overlap. By passing through this area the integration path can enter the target region of $\boldsymbol{x}_2$ without crossing any 'horizontal' contours ($\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_2} > 0$), yielding an unintuitive result. We name this phenomenon "attribution transfer". Though this discussion refers to regions, the same principle applies for smoothly varying functions.

Path B attributes an input vector with both components above the target range. IG explains the zero result as the balance of two competing contributions, unlike BShap and Shap which consider the components to contribute equally. Due to sharp features of the input-output landscape near the integration path, the IG attribution will undergo an abrupt reversal as the input crosses the $\boldsymbol{x}_1 = \boldsymbol{x}_2$ boundary, even though neither the input nor baseline are near a decision boundary, nor in the target range. IG attribution results can be arbitrarily affected by small sharp regions that fall across the path. Is it reasonable to be concerned these effects may occur in practice? We feel it is given that gradients are known to fluctuate sharply (Smilkov et al., 2017), in addition to the results we present.

3.2 AXIOMATIC FOUNDATIONS

Like the other methods, IG is based on Shapley theory and has rigorous axioms. It might then be surprising that IG permits 'maximally misleading' results. IG is based on a variant of Shapley Values known as Aumann-Shapley (Aumann & Shapley, 1974), and has similar but not identical axioms. Both IG and BShap have axioms named *Efficiency*, *Linearity*, *Dummy* (also called *Null Player*), *Affine Scale Invariance*, and *Symmetry*, with IG having additionally *Proportionality* and BShap additionally *Demand Monotonicity* (Sundararajan & Najmi, 2020). However, the Dummy axiom if IG (Sundararajan et al., 2017) is not the Dummy axiom of the original Shapley Values. For clarity we will give separate names to each variant:

Definition 1 (Weak Dummy)**.** An attribution method φ satisfies Weak Dummy iff

$$f(\boldsymbol{z}) = f(\boldsymbol{z} + \alpha \mathbf{1}_i)\ \forall \boldsymbol{z}, \alpha \Rightarrow \varphi(f, \boldsymbol{x}, i) = 0$$

This simply means that if the model does not use an input feature, it has zero attribution.

Definition 2 (Strong Dummy (BShap))**.** Let $\boldsymbol{a}$ be a vector with each component $\boldsymbol{a}_k$ drawn from the corresponding component of either of two given vectors: $\boldsymbol{x}_k$ or $\boldsymbol{x}'_k$, with $\boldsymbol{a}_i = \boldsymbol{x}_i$ for a given i. Let $\boldsymbol{b} = \boldsymbol{a}$ except that $\boldsymbol{b}_i = \boldsymbol{x}'_i$. Then an attribution method φ satisfies Strong Dummy iff $f(\boldsymbol{a}) - f(\boldsymbol{b}) = 0\ \forall \boldsymbol{a} \Rightarrow \varphi(f, \boldsymbol{x}', \boldsymbol{x}, i) = 0$

Strong Dummy is identical to Weak Dummy except that only component values from the target and baseline vector are considered, not the whole range. The weaker conditions make the axiom stronger. Both IG and BShap are formulated to satisfy Weak Dummy, but BShap additionally possesses Strong Dummy. This precludes it from producing unintuitive attributions of the kind given by IG in Table 1. We feel a clear distinction between the axioms in the (original) discrete case and the continuous case may be helpful to avoid any misconception about the strength of the guarantees provided, which are considered extremely robust for Shapley Values in their original context.

4 INTEGRATED CERTAINTY GRADIENTS

We have shown that certain features of the input-output landscape can lead to unintuitive IG attributions. To contrast this behavior in our experiments (section 5) we introduce an IG based method, "Integrated Certainty Gradientss", which does not traverse the original input space. To do this, we append an additional component, the *certainty*, $\boldsymbol{c}$, to the model input, so the model bases its prediction on the input and the *certainty* together: $(\boldsymbol{x}, \boldsymbol{c})$. The dimension of $\boldsymbol{c}$ is the same as the number of attributed features in $\boldsymbol{x}$, and it takes values between $[0, 1]$ that indicate the amount of information the corresponding feature provides. There is a degree of freedom in the choice of features for attribution; for example, in the case of $\boldsymbol{x}$ being an RGB image of size $h \times w$, one can consider each color channel of each pixel to be a feature, in which case $\boldsymbol{c}$ would have a dimension of $h \times w \times 3$. Alternatively one can consider each pixel with all color channels as one feature, giving $\boldsymbol{c}$ dimension $h \times w \times 1$. In our experiments we always take features to be pixels with all color channels combined.

The per-pixel, per-input certainty c is drawn from a random distribution throughout the training of a model, and the model must be trained to associate low certainty values with a concept of *missingness* suitable for the Shapley theory. This is done using a training process that applies a *damage function* D to the input images, so the model receives *damaged inputs*: $(D(\boldsymbol{x}, \boldsymbol{c}), \boldsymbol{c})$. A diagram showing the *damage* process is given in Figure 3. First, *damage data* is required, of the same dimension as the undamaged input $\boldsymbol{x}$. This can take many forms, but for now we will use random noise where each component is i.i.d. uniformly in $[0, 1]$. Then a *damage mask* is generated by treating each component $\boldsymbol{c}_i$ as the probability of keeping the corresponding feature $\boldsymbol{x}_i$. If $\boldsymbol{x}_i$ is not kept it is replaced the corresponding value from the *damage data*.

To use this process in training, a method is needed to generate $\boldsymbol{c}$. Our method will work reasonably well if each component $\boldsymbol{c}_i$ (and thus the keep probability of each 'separate feature') is i.i.d. uniformly in $[0, 1]$. Regular (undamaged) inputs to the model correspond to using $\boldsymbol{c}_i = 1\ \forall\ i$. The features $D(\boldsymbol{x}, \boldsymbol{c})_i$ of an input prepared in this way have varying chances, equal to their certainties $\boldsymbol{c}_i$, to provide information about the original input $\boldsymbol{x}$, and therefore the class $\hat{y}$. During training the model learns to ignore features unrelated to $\hat{y}$, discounting features with lower certainty, and develops the required uncertainty semantics.

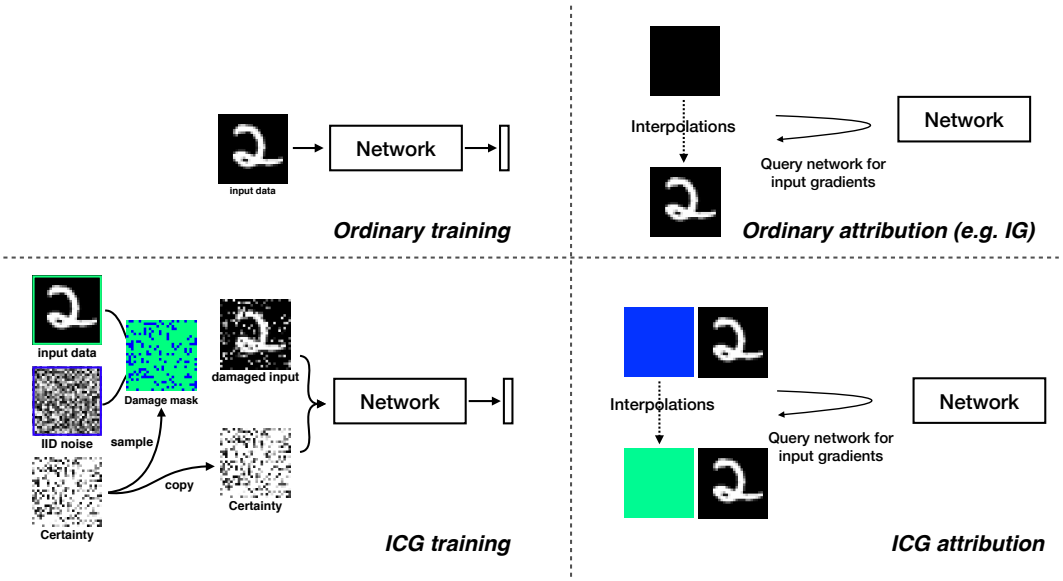


Figure 3: Visualisation of ordinary vs. Integrated Certainty Gradients (ICG) training and attribution. The top row shows ordinary training and attribution using Integrated Gradients which interpolates between a baseline (here: black) and the input example. The bottom row shows our proposed ICG approach. During training, we sample a certainty vector and sample damage using the probabilities from the certainty vector. The model takes this certainty vector as input and thus learns how to account for probabilistically reliable input features. ICG attribution is then simply the Integrated Gradient from no information to full information about an input example.

If the model is trained well, it will ignore features with zero certainty, because these give no information about the original input:

$$\forall\, i : \left.\frac{\partial F\left(\boldsymbol{x}, \boldsymbol{c}\right)}{\partial \boldsymbol{x}_i}\right|_{\boldsymbol{c}_i = 0} \approx 0 \tag{2}$$

In this case, $\boldsymbol{c} = \mathbf{0}$ provides a natural baseline to represent missing data, as the model will ignore the input values entirely.

We introduce the Integrated Certainty Gradients attribution method by combining this baseline with Integrated Gradients. The integral of Equation 1 is taken from $\boldsymbol{c} = \mathbf{0}$ to $\boldsymbol{c} = \mathbf{1}$, holding $\boldsymbol{x}$ constant at the input we wish to attribute. Note that $\boldsymbol{x}$ is *not* damaged while calculating the attribution. It is convenient to note when implementing that the distance term from Equation 1 disappears because $\boldsymbol{c}_i - \boldsymbol{c}'_i = 1$. This integration across certainty space is an Aumann-Shapley version of SHAP in the same way Integrated Gradients is of BShap. Hence we have:

Shapley Values	BShap	SHAP
Aumann–Shapley Values	IG	ICG

Deep networks are capable of detecting and ignoring noise, so the *damage data* described above may not guarantee learning a correct interpretation of certainties. To avoid this, samples from the input dataset can be used as *damage data*, ensuring certainties must be used to distinguish original and damaged features. During ICG the model will encounter interpolations with total mean *certainty* from 0 to 1. Most inputs generated with *certainty* sampled I.I.D. from U(0, 1) will have mean *certainty* close to 0.5. To ensure the model is trained on a wider range of *certainty* inputs we developed a parameterized distribution based on the Continuous Bernoulli (Loaiza-Ganem & Cunningham, 2019) where the mean can be varied betwen 0 and 1. Full information about the training methods used is given in Section A.1.2 of the supporting material.

5 EXPERIMENTS

5.1 BURNT SNACKS SCENARIO

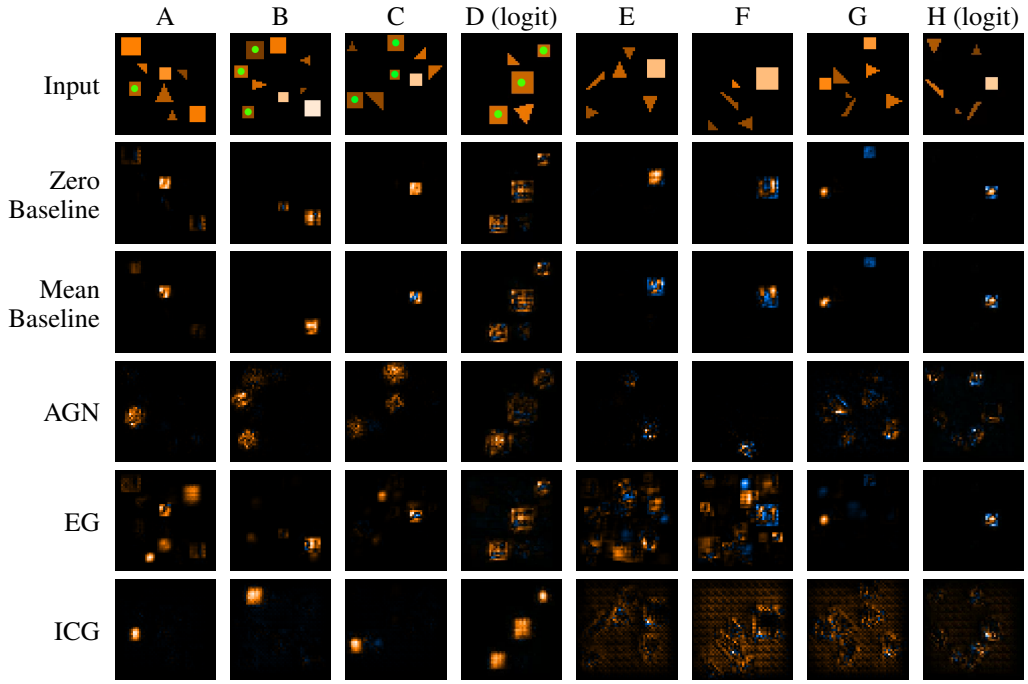


Figure 4: Attribution results for the Burnt Snacks scenario. Each column shows an input image (top row) with associated attributions (other rows). Column A-D inputs are randomly chosen from class 1 (dark square present). Column E-G inputs are randomly chosen from class 0 (no dark square). Probabilities are attributed except columns D and H where logits are attributed. The target objects (dark squares) are marked with a green dot (not present in actual input vector). Expected Gradients is evaluated with 500 interpolations. Additive Gaussian Noise is evaluated with 50 baselines with 100 interpolations each. The other methods are evaluated with 100 interpolations.

We use a purpose designed scenario to demonstrate attribution transfer of methods including Expected Gradients. The scenario is notionally similar to an Optical Sorting task in which overcooked food items are identified (Pedreschi et al., 2016). The model must determine whether 'overcooked' (dark) squares (25% - 50% lightness) are present (class 1), or not (class 0), amidst light squares and other shapes on a dark background. The input-output landscape for an input with two squares with lightnesses x_1, x_2 is equivalent to Figure 2. Two characteristics are designed to mislead Expected Gradients. First, the average pixel value is very dark. This will tend to darken the features of an attributed image when it is interpolated. Second, the light squares are lighter versions of the target objects (dark squares), but are not relevant to the prediction outcome. This combination will cause light squares to often present as dark squares in interpolations, despite having no correlation with them, confusing Expected Gradients. It is easy to justify that light squares should not be attributed, because their presence does not influence the class (relative to black background). A selection of results from a model trained to 98.7% accuracy on this scenario are shown in Figure 4. We attribute probabilities, in line with a prominent guide (Google, 2021), and also logits.

Zero and mean baseline Although a zero baseline seems appropriate, both zero and mean baselines markedly show attribution transfer. Example G for the zero baseline case is equivalent to path B of Figure 2. There are two light squares, one with negative and one positive attribution. By checking the interpolation gradients, we saw the positive attributed square enter the target region (at 25% lightness) first, picking up attribution. The second square followed it but gathered no attribution due to saturation of the output. On leaving (at 50% lightness) the situation was reversed. Attribution transfer also occurs in class 1 scenarios where a light square is present, for example Zero Baseline

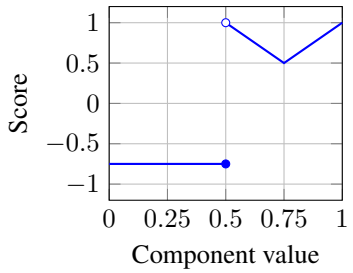


Figure 5: Component-wise score function f for the Local Minimum scenario. The total score is the sum of the scores of the covariates $S(\boldsymbol{x}) = \sum_i f(\boldsymbol{x}_i)$. Component values range from 0 to 1. The score is -0.75 for component values 0.5 and below, and between 0.5 and 1 for values above 0.5.

example C. This situation is equivalent to path A of Figure 2. As a result, the process concludes with only the light square, irrelevant to the prediction, attributed.

Expected Gradients Results for Expected Gradients are shown in the 5th column of Figure 4. In no case does EG identify the dark squares when they are present when using probabilities (class 1, images A-C). Instead, it transfers attribution to the light squares, most strongly the lightest. In case A it appears no square is light enough to attract attribution, and artifacts from the random sampling process are shown.

Since class 0 is an absence class the expected outcome is not as clear. Images E and F show diffuse artifacts from the interpolation process and arguably this is a fair result. However image G strongly attributes a specific square and this seems more clearly misleading, as the square has no significance to the classification outcome.

Integrated Certainty Gradients ICG does not suffer from the previously described problems. Because the integration is across the uncertainty space, the relative location of features in the input space and the input distribution are not important. For class 0, ICG attributions are very diffuse and largely uniform, which could be considered an appropriate outcome, arguably the most intuitive of any method. For class 1, it always attributes a dark square (correct target feature). The attributions are strong and precisely located. However, only one square is attributed even when multiple are present. In each case the darkest square is attributed. Removing the darkest square results in a similar clear attribution for the second darkest. Repeating the process attributes the remaining dark square. The reason for this behavior is not clear. We posit that it may be reflecting a model behavior; perhaps the model first identifies the darkest square in the image, then tests whether that square falls within the target range. This explanation cannot fully reconcile the behavior with well trained uncertainty awareness. Further investigation into this behavior would likely be beneficial.

Additive Gaussian Noise Additive Gaussian Noise correctly attributes the location of the dark squares in class 1 images. The attributions are noisy and not as precisely localised to the squares as those of ICG, however, all squares are attributed. In class 0 attribution focuses around the triangles. This is unsurprising because the addition of noise could easily make them resemble squares.

Logits For most methods logits show intermixed opposite attributions on both light and dark squares. Dark squares have overall more positive attribution, but nonetheless this presentation is imperfect and potentially confusing. ICG performs excellently in this scenario, no longer missing squares, attributing all dark squares clearly.

5.2 Local Minimum scenario

Additive Gaussian Noise uses baselines drawn mainly from a region local to the input vector. We introduce a scenario designed to challenge local baselines. Input vectors are classified according to a total score $S(\boldsymbol{x})$, which is the sum of component-wise scores $f(\boldsymbol{x}_i)$. Inputs are in class 1 if $S(\boldsymbol{x}) > 0$, class 0 otherwise. Note that although S is an additive function, the classifier as a whole is not. The component-wise score function f is shown in Figure 5. f has a positive local minimum. This local minimum makes a positive contribution to class 1 even though it is less contributing than nearby values. Attribution methods based on local baselines would be expected to give the local minimum

Table 2: Comparison of attribution methods on the Local Minimum scenario. Each row corresponds to a component of the input vector. The first column shows the input vector. The second column shows the "scores" of the components contributing to the true class, normalised to maximum value 1, for comparison with attribution results. The remaining columns show attribution results, normalised to maximum value 1, from BShap, zero baseline Integrated Gradients, mean baseline Integrated Gradients, Expected Gradients, Additive Gaussian Noise and Integrated Certainty Gradients. Attributions are shown for class 1, not the predicted class (0), to aid comparison with the score values.

Input vector	Scaled nominal score	BShap	Zero IG	Mean IG	EG	AGN	ICG
0.1	-0.94	-0.75	0.001	-0.01	-0.30	0.004	-0.75
0.2	-0.94	-0.74	0.002	-0.001	-0.73	-0.004	-0.75
0.3	-0.94	-0.74	0.003	-0.37	-0.66	0.004	-0.74
0.6	1	1	1	1	0.82	1	1
0.7	0.75	0.50	-0.24	0.89	0.70	-0.42	0.48
0.9	1	0.76	0.23	-0.07	1	-0.02	0.74

an incorrect negative attribution. We trained a certainty-aware network on this classification scenario to 93% accuracy. Attribution results for this network using an example input vector from class 0 are shown in Table 2. Component values of 0.50546 were used for the BShap baseline, as this was determined to be close to the model's prediction boundary.

BShap, ICG and EG performed well in the scenario, with ICG and BShap in excellent agreement. Only IG and BShap accurately represent the uniform negative region. All other methods performed badly. As expected, AGN incorrectly gives the local minimum a negative attribution. Zero IG shows strong signs of attribution transfer. Zero IG and AGN both exhibit blindness in the negative score region due to sharing this negligible gradient region with the baselines. The results for mean IG vary widely. The shown result is a randomly chosen example. The distribution mean lies on the discontinuity of the score function (0.5) so small variations around it make major differences to the result.

6 Conclusion

We have described a situation in which Integrated Gradients produces extremely unintuitive results, and explained how this arises from areas where the prediction function is non-additive ("attribution transfer"). This behavior may be unexpected because IG is founded in the theory of Shapley Values, an axiomatic framework for fair allocation. We explain this seeming contradiction by identifying a difference in the formulation of an axiom also present in the original theory.

In experiments we have shown that attribution transfer occurs even for state of the art methods like Expected Gradients in the presence of a highly skewed but plausible data distribution. Avoiding detrimental interactions by restricting IG to a small area of the input domain, as Additive Gaussian Noise does, does not prevent misleading results when the local prediction landscape does not reflect the wider input space.

We show these failures of IG may be avoidable, by developing a method with a new kind of baseline, Integrated Certainty Gradients, that requires a special training process to learn the data distribution. This method is the Aumann-Shapley approximation to SHAP (as IG is to BShap) and by not traversing the input space it avoids the failure cases we demonstrate in other methods.

Acknowledgments

LEFT BLANK IN ANONYMOUS VERSION

ETHICS STATEMENT

Attribution is of key importance in many scenarios with ethical implications, for example providing reasons for life affecting decisions such as loan assessments, or determining whether decisions are made based on legitimate or protected personal characteristics. As such, the development of fair attribution methods is ethically important and this paper has been produced with appreciation of that. The work itself undertaken to produce this paper did not raise any ethical concerns.

REPRODUCABILITY STATEMENT

Model architectures and training hyperparameters are given in Appendix A.2. Anonymized source code is available at https://github.com/account48294/iclr1

REFERENCES

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

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020.

R. J. Aumann and L. S. Shapley. *Part I: The Axiomatic Approach*, pp. 11–92. Princeton University Press, 1974. URL http://www.jstor.org/stable/j.ctt13x149m.5.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 648–657, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375624. URL https://doi.org/10.1145/3351095.3375624.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation, 2020.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical report, University of Montreal, 2009.

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients, 2020.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 2017. doi: 10.1109/iccv.2017.371. URL http://dx.doi.org/10.1109/ICCV.2017.371.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, October 2017. ISSN 0738-4602. doi: 10.1609/aimag.v38i3.2741. URL http://dx.doi.org/10.1609/aimag.v38i3.2741.

Google. Integrated gradients, 2021. URL https://www.tensorflow.org/tutorials/interpretability/integrated_gradients.

Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. On baselines for local feature attributions, 2021.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Gabriel Loaiza-Ganem and John P. Cunningham. The continuous bernoulli: fixing a pervasive error in variational autoencoders, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

F. Pedreschi, D. Mery, and T. Marique. Chapter 22 - quality evaluation and control of potato chips. In Da-Wen Sun (ed.), *Computer Vision Technology for Food Quality Evaluation (Second Edition)*, pp. 591–613. Academic Press, San Diego, second edition edition, 2016. ISBN 978-0-12-802232-0. doi: https://doi.org/10.1016/B978-0-12-802232-0.00022-0. URL https://www.sciencedirect.com/science/article/pii/B9780128022320000220.

Lloyd S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. doi: 10.7249/P0295.

K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. pp. 1–8. ICLR, 2014.

Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 2020. ISSN 2313-433X. doi: 10.3390/jimaging6060052. URL https://www.mdpi.com/2313-433X/6/6/52.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, July 2020. URL http://proceedings.mlr.press/v119/sundararajan20b.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Gradients of counterfactuals, 2016.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance, 2019.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.

A Appendix

A.1 Training

A.1.1 Validating training

For effective training, methods to assess correct interpretation of certainty are required. We introduce two techniques for this.

Information absence To ensure fair attributions, the model should interpret all baselines in the same way, as formalized by equation 2. This may be verified by varying $\boldsymbol{x}$ with $\boldsymbol{c} = \mathbf{0}$ and confirming predictions do not change.

Integrated Gradients gives zero attribution when components match the baseline (baseline blindness). However, Integrated Certainty Gradients relies on the model itself also ignoring zero *certainty* pixels, so that model and attribution behavior are aligned. This means that zero certainty pixels should not accrue attribution by any method, not just ICG. This makes $\boldsymbol{c} = \mathbf{0}$ in a sense a universal baseline. This motivates a distinctive visual test. If the model is well trained, it should be possible to suppress the attribution of an accurate method to a negligible amount by masking out that area with zero certainty during the attribution process. This can be tested by using IG with baselines $\boldsymbol{x}' = \mathbf{0}$ and $\boldsymbol{x}' = \mathbf{1}$: the result should show an obvious suppression of attribution in the uncertain region in both cases. This test is visualized in Figure 6.

Neutrality Zero *certainty* baselines should provide no input information to the model, so its output should be unbiased. It is therefore reasonable to expect the model prediction in the baseline space to match its expectation on the training set: $F(\boldsymbol{x}, \boldsymbol{c} = \mathbf{0}) \approx \mathbb{E}(F(X_{\text{train}})) \; \forall \; \boldsymbol{x}$.

The expectation of each class in the training distribution is the proportion of samples in that class. So for example in one hot representation, $\hat{\boldsymbol{y}}_i$ should be approximately the proportion of the training distribution occupied by class i.

Using these tests we assessed a number of methods for certainty damage. The following methods improved training compared to using uniform distributions.

A.1.2 DAMAGE FUNCTIONS

The basic *damage function* described in section 4 can produce informative attributions but more effective *damage functions* have been identified.

Variable damage extent With uniform random *certainty*, the trained model often widely violates the neutrality condition. Due to regression towards the mean, i.i.d. uniform random *certainty* results in an image with mean *certainty* $\frac{1}{n}\sum_i \boldsymbol{c}_i$ typically close to 0.5. It is hypothesized that the the neutrality condition is not learned due to the lack of training inputs with mean *certainty* near zero. We define *damage extent* as the mean uncertainty of an input: $1 - \frac{1}{n}\sum_i \boldsymbol{c}_i$. A *damage function* which results in *damage extents* distributed uniformly in [0, 1] will expose the model to inputs along more of the *certainty* interpolation space.

Sampling from the continuous Bernoulli distributionLoaiza-Ganem & Cunningham (2019) was chosen as a new *certainty* generating function because it produces samples in the range [0, 1] and has a parameter which allows the expectation to be varied in [0, 1] as well as other desirable properties. The parameter does not vary the expectation uniformly but instead concentrates it towards the center of the range. To counteract this we reparameterized the distribution. Let $e(\lambda)$ be the expectation of the distribution with parameter λ. Then $\lambda' := e^{-1}(\lambda)$, where e^{-1} is calculated numerically from the analytical form of e. Using this new parameter gives a distribution parameterized by the mean. It was found that varying *damage extent* using this distribution does in fact improve neutrality.

Adversarial damage Random noise is typically not strongly biased towards an input class. Uniform random *damage data* therefore results in an input which is a combination of data from the original class and unbiased data. As a result, the input is always positively biased towards the original class. Furthermore, if the model can detect and exclude random noise, it will be able to determine the true class from the remaining data. These considerations suggest the model might be able to ignore $\boldsymbol{c}$ and still achieve good predictions.

Motivated by this, *adversarial damage* was tested: using a random alternative sample from the test dataset as the *damage data*. When the alternative and original input are from different classes, damaged features $D(\boldsymbol{x}, \boldsymbol{c})_i$ provide evidence for a different class from the true one and with a large *damage extent* the majority of features can indicate the alternative class. It is not possible for the model to make an accurate prediction in such circumstances without considering $\boldsymbol{c}$, forcing the model to use it. *Adversarial damage* was found to train certainty awareness faster and with more accuracy

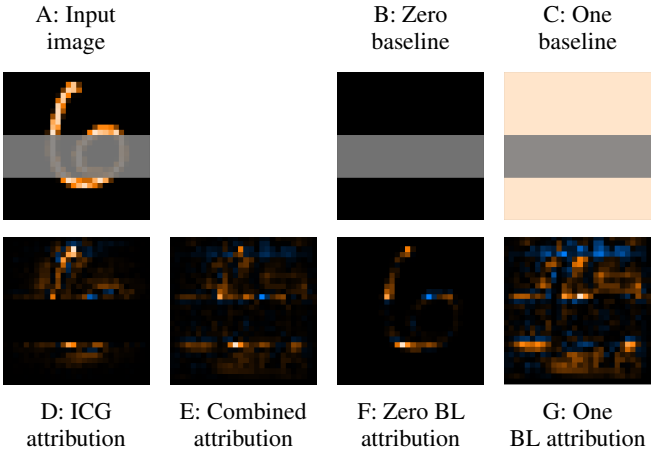


Figure 6: Example of the Information Absence test. An area of the image is masked out to zero *certainty* during attribution (shown in grey). Integrated Gradients are taken from uniform zero and uniform one baselines. Even though these are not certainty-aware attribution methods, they show the same exclusion of attribution from zero certainty areas as ICG, suggesting that zero *certainty* is a "universal" baseline. Image E shows the attributions for the zero baseline and one baseline added together.

than random uniform damage. Another consideration is that it avoids exposing the model to out of distribution feature values, although the implications of this were not investigated.

Binary uncertainty Continuous Bernoulli and uniform damage distributions have been described above. Discrete Bernoulli damage was also tested, producing $\boldsymbol{c}_i$ values in $\{0, 1\}$. It was found that this improves learning rate compared to uniform random damage. The reason for this has not been determined, although it may be since $\boldsymbol{c}_i$ can be only 0 or 1, the value components no longer require a probabilistic interpretation, which makes the scenario easier for the model to learn.

We found combining adversarial and uniform *damage data*, binary and uniform *uncertainty*, and using variable *damage extents* resulted in the most effective training in terms of speed and final evaluation. This training was used for all results which are discussed below.

A.1.3 Learning rate and accuracy

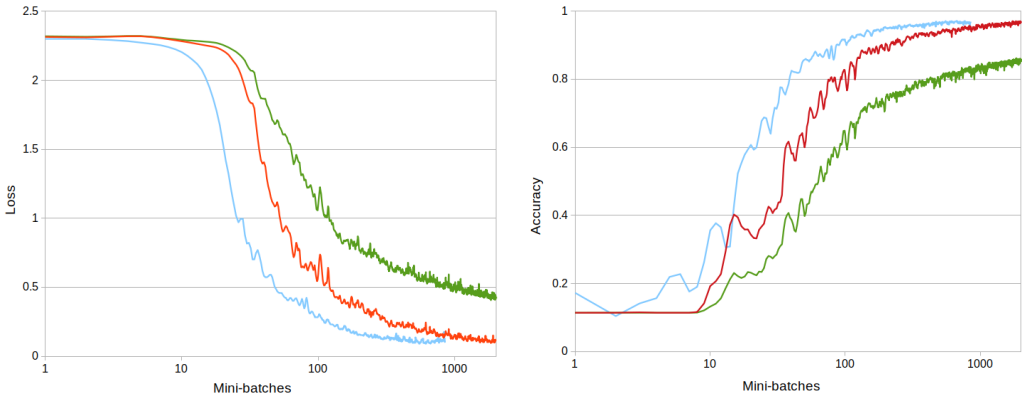
Certainty awareness increases the classification task complexity. To assess the impact of this we trained MNIST models with and without input damage. The results are shown in Figure 7. The red and blue curves show training with and without damage respectively. Both are evaluated on undamaged inputs. They show that damage training has not resulted in reduced accuracy on MNIST for this model, although training takes twice as many batches. Variance in accuracy and loss is also increased. Given that on average half the components of the original image are lost during damage, this seems a modest penalty. The green curve shows results for evaluation of the damage trained model on damaged inputs. This is included for completeness but not especially relevant to the comparison.

A.2 Experimental details

Model architectures are shown in Figures 10, 8, and 9.

The *damage function* D used during training is specified by a combination of *damage data* and *certainty* generator. For all models *damage data* was taken from either $U(0, 1)$ or *adversarially* from another training input. Within each input, $\boldsymbol{c}_i$ values were taken i.i.d from either the Bernoulli or our reparameterization of the continuous Bernoulli distribution. Both have a parameter equal to their expectation, and for each image this was sampled from $U(0, 1)$. The four combinations of *damage data* selection method and *certainty* generation method were each used in equal proportion.

Xavier initialization was used for network weights and network biases were initialized to zero.



A: Training loss

B: Training accuracy

Figure 7: Training progress with and without damaged inputs. Blue curves show learning progress without damage. Red curves show learning progress with damage evaluated on undamaged inputs. Green curves show learning progress with damage evaluated on damaged inputs. The evaluation images were drawn from the test dataset. Progress is measured in batches (of size 64) because multiple damaged inputs were created from each original input, leading to differently sized epochs. Evaluations on damaged inputs are significantly less accurate, which is to be expected given that when the *damage extent* is high they may contain little of the original image. Evaluations of the damage trained model on undamaged inputs reach a similar accuracy to the model trained without damage. Their curves show a roughly constant horizontal offset. On these logarithmic scales this indicates an approximately constant slowdown ratio: training with damage takes roughly twice as long.

Burnt snacks Training data was generated procedurally hence there is no epoch size or train-test split. Adam optimization was used with learning rate 0.001 and batch size of 64.

Local minimum Training data was generated procedurally hence there is no epoch size or train-test split. The network did not converge for several training attempts, and was reinitialized several times until an initial state which did converge was found. The network was initially trained with Adam optimization and a learning rate of 0.001. After initial training, reduction in the learning was found to improve accuracy, and training was continued with reduced learning rates down to a minimum of 10^{-6}. A batch size of 128 was used for initial training and a batch size of 256 for the reduced learning rate stage.

MNIST The dataset was split into 60,000 training examples and 10,000 test examples across all classes. Adam optimization was used for training with learning rate 0.001 and batch size of 64.

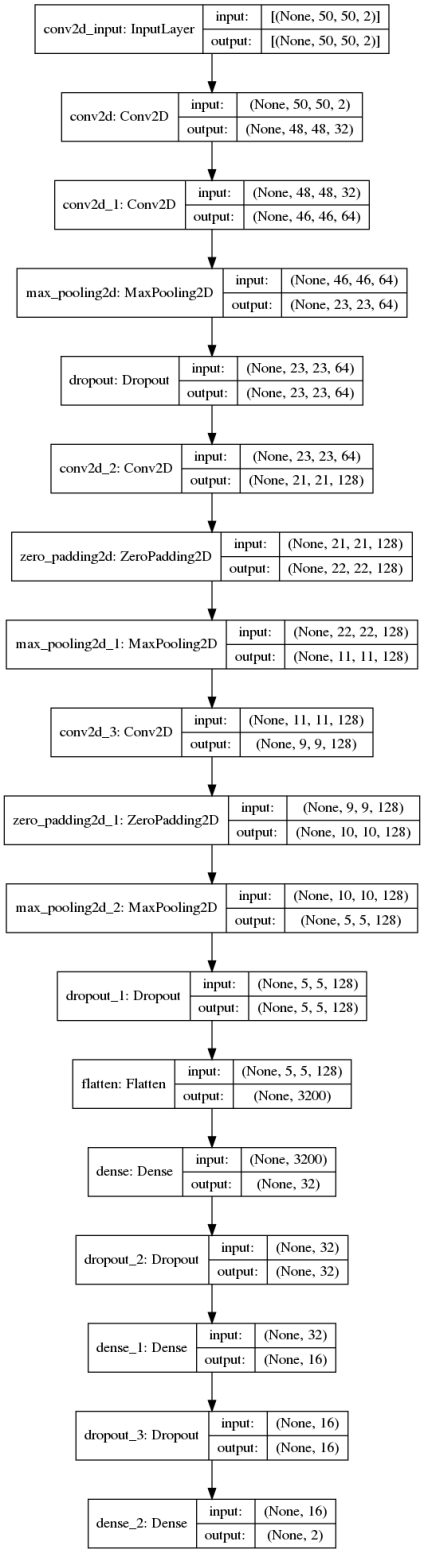


Figure 8: Model architecture used for Burnt Snacks scenario comparison.

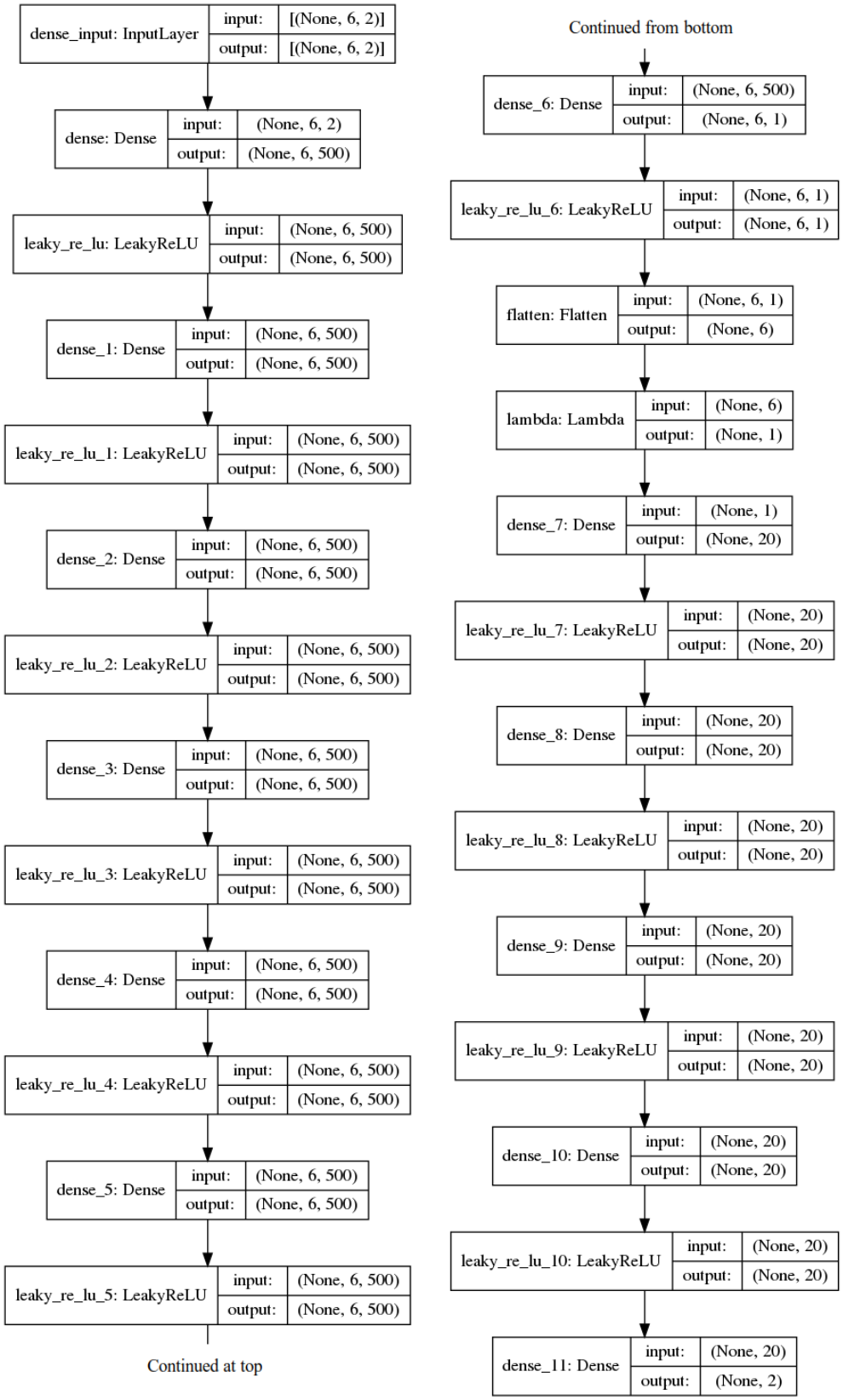


Figure 9: Model architecture used for Local Minimum scenario comparison. The layer 'lambda: Lambda' sums the input components.

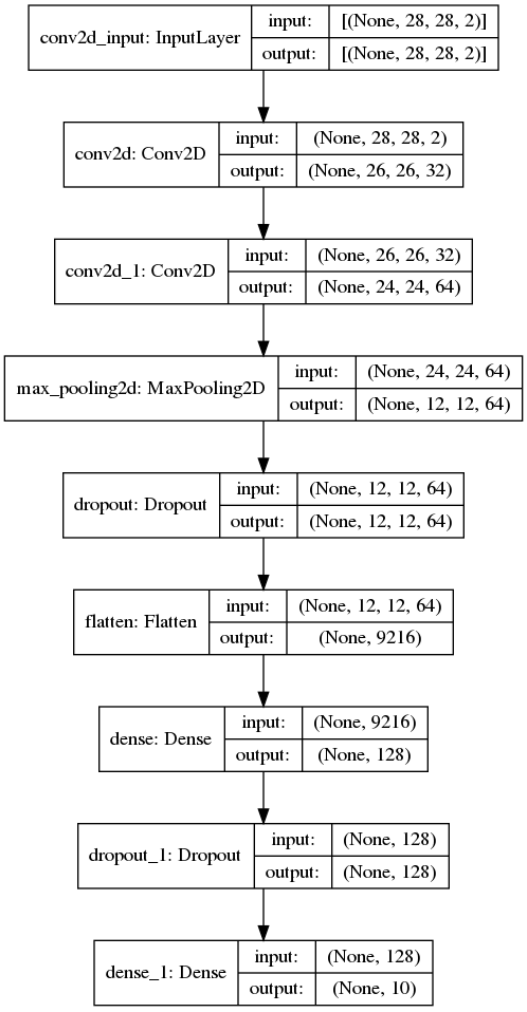


Figure 10: Model architecture used for MNIST testing.