BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations

Xingyu Zhao 1,2, Wei Huang 1, Xiaowei Huang 1, Valentin Robu 2,3,4, David Flynn 2

1Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, U.K.
2School of Engineering & Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, U.K.
3Centrum voor Wiskunde en Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
4Delft University of Technology, Algorithmics Group, EEMCS, 2628 XE Delft, The Netherlands

Abstract

Given the pressing need for assuring algorithmic transparency, Explainable AI (XAI) has emerged as one of the key areas of AI research. In this paper, we develop a novel Bayesian extension to the LIME framework, one of the most widely used approaches in XAI – which we call BayLIME. Compared to LIME, BayLIME exploits prior knowledge and Bayesian reasoning to improve both the consistency in repeated explanations of a single prediction and the robustness to kernel settings. BayLIME also exhibits better explanation fidelity than the state-of-the-art (LIME, SHAP and Grad-CAM) by its ability to integrate prior knowledge from, e.g., a variety of other XAI techniques, as well as verification and validation (V&V) methods. We demonstrate the desirable properties of BayLIME through both theoretical analysis and extensive experiments.

1 INTRODUCTION

A key challenge to wide adoption of AI methods is their perceived lack of transparency. The black-box nature of many AI methods means they do not provide human users with direct explanations of their predictions. This led to growing interest in Explainable AI (XAI) – a research field that aims at improving the trust and transparency of AI. Explainable AI (XAI) methods can be classified by various criteria [Molnar, 2020, Chpt. 2.2], such as model-specific vs model-agnostic or local (instance-wise) vs global (entire model). Readers are referred to [Huang et al., 2020, Adadi and Berrada, 2018] for a comprehensive review.

In this work, we focus on the class of XAI methods using local surrogate models for explaining individual predictions – specifically, we develop a novel Bayesian extension to the most popular method in this category: Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016]. Despite its very considerable success in both research and practice, LIME has several weaknesses, the most significant of which are the lack of consistency in repeated explanations of a single prediction and robustness to kernel settings. Meanwhile, higher explanation fidelity is also expected in many settings. Arguably, these three properties are among the most desirable for an XAI method to have.

The inconsistency of LIME, where different explanations can be generated for the same prediction, has been identified as a critical issue by several prior works [Rehman Zafar and Mefraz Khan, 2019, Shankaranarayana and Runje, 2019]. This is caused by the randomness in generating perturbed samples (of the instance under explanation) that are used for the training of local surrogate models. Plainly stated, a smaller sample size leads to greater uncertainty of such randomness. This inconsistency has been argued to limit its usefulness in critical applications such as the medical domain, where consistency is highly required [Rehman Zafar and Mefraz Khan, 2019]. As an example, in Fig. 1, the pre-trained Convolutional Neural Network (CNN) InceptionV3 [Szegedy et al., 2016] predicts the instance of Fig. 1 (A) as “Bernese mountain dog” (top-1 label). To explain this prediction, we vary the size of perturbed samples (denoted as \( n \)) and record the time consumption in Fig. 1 (B). We see that the computational time is linear with respect to \( n \). If, say, an application requires LIME to respond in 20s, we have to limit \( n \) to around 100 (in our case). Then, we may easily get three inconsistent explanations, as shown in Fig. 1 (C)-(E), in three repeated runs of LIME.

The other problem of LIME motivating this work relates to defining the right locality – the “neighbourhood” on which a local surrogate is trained to approximate the AI model [Laugel et al., 2018]. As noted by [Molnar, 2020, Chpt. 5.7] as a “very big and unsolved problem”, how to choose a kernel setting in LIME to properly define a neighbourhood is challenging. Even worse, explanations can be easily turned around by changing kernel settings (cf. the example in [Molnar, 2020, Chpt. 5.7]). Thus, an enhancement to LIME so
that it is robust to the kernel settings is called for.

The goal of XAI is to provide good explanations. Despite the various definitions on the goodness of explanations, fidelity – how truthfully the explanation represents the underlying AI decision – is the most important criterion. Any new XAI method enhancing other properties should preserve the fidelity (if not improving), compared to the state-of-the-art.

In this paper, we propose a novel enhancement of LIME, named as BayLIME, by utilising a Bayesian local surrogate model, which we show analytically is a “Bayesian principled weighted sum” of the prior knowledge and the estimates based on new samples (as LIME). The weights consist of parameters with dedicated meanings that can be either automatically fitted from samples (Bayesian model selection) or elicited from application-specific prior knowledge (e.g., V&V methods). Our experiments show that BayLIME significantly improves over LIME with respect to a few objective measures, because the prior knowledge is independent from the causes of aforementioned inconsistency and unrobustness (thus benefits both properties) and also includes additional useful information that improves the fidelity.

The contributions of our work include:

(1) Developing a new XAI method, BayLIME, that provides a Bayesian principled mechanism for the embedding of prior knowledge;

(2) Introducing and designing quantitative metrics on both the consistency in repeated explanations and robustness to kernel settings that are generic to all feature-ranking based XAI techniques;

(3) Three typical ways of obtaining prior knowledge and eliciting prior parameters, including explanations on similar instances, other diverse XAI methods and V&V methods, to illustrate potential use cases of BayLIME;

(4) An extensive set of experiments, exhibiting the superior performance of BayLIME over LIME on aforementioned consistency and robustness, and higher fidelity over the state-of-the-art (LIME, SHAP, GradCAM);

(5) A prototype tool with tutorials/experimental-results.

2 CONSISTENCY IN REPEATED EXPLANATIONS

LIME is known to be unstable in terms of generating inconsistent explanations in repeated runs [Rehman Zafar and Mezraz Khan, 2019, Shankaranarayana and Runje, 2019]. This is due to the randomness in the perturbed samples used to train the local surrogate model. While randomness can be reduced by enlarging the sample size, it might be impractical in real-world applications, particularly those applications with strict efficiency constraints (e.g., low response time).

As shown in Fig. 1 (B), to improve LIME’s efficiency, we cannot effectively tune other LIME arguments, rather the best option is to limit the number of queries made to the AI model (which, indeed, is very costly when the AI model is deep [He and Sun, 2015]) when generating labels for the $n$ samples (step 3 of the LIME workflow in Appendix C). However, this solution leads to a severer inconsistency issue.

To compare between BayLIME and LIME on their inconsistency, we take Kendall’s W [Kendall and Smith, 1939] to measure the agreement among raters (i.e. repeated explanations in our case), which ranges from 0 (no agreement) to 1 (complete agreement). However, Kendall’s W only considers the discrete ranks of features, and cannot discriminate explanations with the same ranking of features but different importance vectors. To complement Kendall’s W in such corner cases, we also introduce a new metric based on the index of dispersion (IoD) of each feature in repeated runs. The new metric weights the IoD of the rank of each feature by its importance, cf. Appendix B for a formal definition.

3 ROBUSTNESS TO KERNEL SETTINGS

Another notorious problem of LIME (or any approach with a notion of localisation) is how to meaningfully define a “neighbourhood” of the instance of interest (as required by the step 4 of LIME depicted in Appendix C). By default,
LIME uses an exponential smoothing kernel to define the neighbourhood with a given kernel width that determines how large the neighbourhood is. While intuitively a smaller kernel width means that an instance needs to be closer to influence the local model (vice versa), no effective way exists to find the best kernel settings. The best strategy for now is to try different kernel settings and see if the explanations make sense, which inevitably is subject to errors/bias. Moreover, an explanation may be easily turned around by changing the kernel settings [Molnar 2020 Chpt. 5.7].

To compare between BayLIME and LIME, we explore the robustness to the kernel width parameter (denoted by \( l \) later) in LIME’s default exponential kernel function. Other kernel setting parameters can be studied in a similar way.

**Definition 1** To explain a given instance \( i \), we denote \( h_i : (0, +\infty) \rightarrow \mathbb{R}^m \) as the importance vector of \( m \) features taking \( l \) as the kernel width setting. For any pair of kernel width parameters \( l_1 \) and \( l_2 \), there exists a global Lipschitz value \( L \in \mathbb{R} \) such that ||\( h_i(l_1) - h_i(l_2) \)|| \leq L||l_1 - l_2||.

Intuitively, this global Lipschitz value quantifies the robustness of an explainer to the change of kernel width. The computation of \( L \) can be seen as an optimisation problem:

\[
L = \arg \max_{l_1, l_2 \in (0, +\infty)} \left\{ \frac{||h_i(l_1) - h_i(l_2)||}{||l_1 - l_2||} \right\}
\]

which, unfortunately, is very challenging to solve analytically or to estimate numerically, due to the high complexity and non-linearity of the function \( h_i \). Similar difficulties are also in, e.g., [Alvarez-Melis and Jaakkola 2018], when studying the robustness to perturbed input instances.

To bypass the difficulty and still provide insights on the robustness defined earlier, we instead introduce a weaker empirical notion of the robustness to kernel settings:

**Definition 2** Assume \( L_1 \) and \( L_2 \) are both random variables of kernel width settings within \( [l_{lo}, l_{up}] \). Then, \( R \) is the median (denoted as \( \mathbb{M}() \)) of the ratio between the perturbed distances of \( h_i \)s and the pair \( (L_1, L_2) \):

\[
R = \mathbb{M}\left( \frac{||h_i(L_1) - h_i(L_2)||}{||L_1 - L_2||} \right)_{l_{lo} \leq L_1 \leq l_{up}, l_{lo} \leq L_2 \leq l_{up}}
\]

which represents the average robustness to the kernel settings when explaining the instance \( i \).

The kernel width cannot be too large nor too small (cf. the general discussions on variance-bias trade-off in selecting hyper-parameters [Cawley and Talbot 2010]). In practice, after a few trials, e.g., by cross-validation, it is not hard to figure out empirically a bound \( [l_{lo}, l_{up}] \) as the range of all possible kernel settings in which we may perturb the kernel width to compute the robustness. In Eq. (2), we focus on the median value which is a much easier quantity to estimate than the optimised solution in Eq. (1), yet it provides insights on the general robustness.

## 4 EXPLANATION FIDELITY

Among many evaluation criteria on the goodness of explanations, fidelity – how truthfully the explanation represents the unknown behaviour of the underlying AI model – is of great importance. In other words, we want the XAI method to explain the true cause of the underlying model’s prediction [Petsiuk et al., 2018; Sun et al., 2020]. In this paper, we adopt the true causality as an indicator for the explanation fidelity in the following two different ways.

First, we reuse the two causal metrics proposed in [Petsiuk et al., 2018] – deletion and insertion. The intuition behind deletion is that the removal of the “cause” (important features) will force the underlying AI model to change its decision. To be exact, starting with a complete instance and then gradually removing top-important features (according to the ranked list obtained from XAI methods), the deletion metric measures the decrease in the probability of the predicted label. A sharp drop and thus a low Area Under the Probability Curve (AUC), as a function of the fraction of removed features, suggests a good explanation that captures the real causality. Similarly, but in a complementary way, the insertion metric measures the increase in the probability as more and more important features are introduced, with higher AUC indicating a better explanation. The two metrics not only alleviate the need for large-scale human annotation effort, but also are better at assessing causal explanations by avoiding human bias [Petsiuk et al., 2018].

Second, we evaluate explanation fidelity through neural backdoors, inspired by [Lin et al., 2020; Sun et al., 2020]. The major difficulty of evaluating fidelity is the lack of ground truth behaviour of the underlying AI model. As the most important features that cause deliberate misclassification, backdoor triggers provide such ground truth and therefore should be highlighted by a good XAI method. Both [Lin et al., 2020] and [Sun et al., 2020] use the metric Intersection over Union (IoU) to measure the success in highlighting the trigger. Given a bounding box around the true trigger area \( B_T \) and the highlighted area by XAI methods \( B'_T \), the IoU is their overlapped area divided by the area of their union, i.e., \( \frac{|B_T \cap B'_T|}{|B_T \cup B'_T|} \). IoU is an established metric originally designed for object detection that ranges from 0 to 1, a higher IoU is better. It only considers the overlapping of the highlighted area and the ground truth, ignoring how geometrically closed they are when there is no overlapping. To complement this, we introduce a secondary
metric in Appendix A.2 to further confirm our conclusions.

5 A BAYESIAN RETROFIT OF LIME

Denote the input set with $n$ samples as $\mathbf{X} = \{x_1, \ldots, x_n\}$ where $x_i$ is the instance $i$ with $m$ features (i.e. $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times m}$), and let the corresponding $n$ target values be a column vector $\mathbf{y} = [y_1, \ldots, y_n]^T$. In Bayesian linear regression, a response $y_i$ is assumed to be Gaussian distributed around $x_i \beta$:

$$Pr(y_i \mid \beta, x_i, \alpha) = \mathcal{N}(y_i \mid x_i \beta, \alpha^{-1})$$

(3)

where $\beta$ is the coefficient vector of $m$ features, and $\alpha$ is the precision parameter (reciprocal of the variance) representing their respective proximity to the instance. There-fore, the posterior estimate on $\mu_n$ of the “pseudo-count” of prior sample size based on which we form our prior estimates $\mu_0$; (ii) $\alpha \mathbf{X}^T \mathbf{X}$, the “accurate-actual-count” of observation sample size, i.e. the actual observation of the $n$ new samples $\mathbf{X}^T \mathbf{X}$ scaled by the precision $\alpha$.

To see the above insight clearer, we present the special case of a single feature instance ($m = 1$) with a simplified kernel function that returns a constant weight $w_i$ (i.e. $w_i = w_c, \forall i = 1, \ldots, n$), then $\mu_n (\mu_0$ with 1 feature) becomes:

$$\frac{\lambda}{\lambda + \alpha w_c \sum_{i=1}^{n} x_i^2} \mu_0 + \frac{\alpha w_c \sum_{i=1}^{n} x_i^2}{\lambda + \alpha w_c \sum_{i=1}^{n} x_i^2} \beta_{MLE}$$

(10)

where $\beta_{MLE} = \frac{\alpha w_c \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. Now, denote $X$ as the r.v. representing $x_i$, we know:

$$\sum_{i=1}^{n} x_i^2 \approx n \text{Var}(X) = n \left( \text{Var}(X) + \text{E}(X)^2 \right)$$

(11)

As implemented by Lime, $x_i$s are $n$ random samples from some distribution, depending on the type of the feature, e.g.: (i) For superpixels of images, $x_i \in \{0, 1\}$ are random samples from an uniform 0-1 two-point distribution; (ii) For numerical features of a tabular dataset, $x_i$ are random samples from a standard Gaussian $\mathcal{N}(0, 1)$. Thus, say, for numerical features from tabular data: $\sum_{i=1}^{n} x_i^2 \approx n (1 + 0^2) = n$ via Eq. (11). Then, Eq. (10) can be simplified as:

$$\frac{\lambda}{\lambda + \alpha w_c n} \mu_0 + \frac{\alpha w_c n}{\lambda + \alpha w_c n} \beta_{MLE}$$

(12)

which neatly shows how the prior knowledge is embedded in our BayLIME in an intuitive story – “based on $\lambda$ data-points prior to our new experiment, we form our prior estimate of $\mu_0$.

Now, in the experiments, we collect $n$ samples. After considering the precision ($\alpha$) and weights ($w_c$) of the new samples, we form a MLE estimate $\beta_{MLE}$. Then, we combine the two estimates $- \mu_0$ and $\beta_{MLE}$ according to their proportions of the effective samples size used, i.e. $\lambda + \alpha w_c n$ (the posterior precision)”. Now it is easy to see via Eq. (12) (also holds for Eq. (9)):

Remark 1 Smaller $\lambda$ means lower confidence in the prior knowledge, thus the posteriors are mostly dominated by the new observations, and vice versa. That is, in extreme cases: 1) When $\lambda \approx 0$ the result (9) (and its simplified case (12)) reduces to MLE, i.e. “let the data speak for themselves”. 2) On the other hand, if $n \approx 0$ (or equivalently $\alpha \approx 0$, $w_c \approx 0$), then the $\beta_{MLE}$ estimate from the new data is negligible and the prior knowledge dominates the posteriors.
This kind of intuitive decomposition is generic to Bayesian inference in the linearly updated canonical-exponential families [Bernardo and Smith [1994], which is found to be useful in applications that exploit prior knowledge, e.g., [Filieri et al., 2012; Zhao et al., 2019; Walter and Augustin, 2009; Zhao et al., 2020].

Remark 2 Via the “weighted sum” Eq. (9) (and its simplified case (12)), the general reasons why integrating prior knowledge improve the three aforementioned properties are: 1) $\beta_{MLE}$ is a function of the $n$ randomly perturbed samples that causes inconsistency, while $\mu_0$ is independent from the cause. A weighted sum of both may “average out” the randomness, thus improving the consistency.
2) $\beta_{MLE}$ is a function of $W$ that depends on the choice of kernel settings, while $\mu_0$ is independent from kernel settings. A weighted sum of both may “average out” the effect from kernels, thus improving the robustness.
3) $\mu_0$ normally contains added diverse information to $\beta_{MLE}$ (black-box queries) that benefits the explanation fidelity.

5.2 THE BAYLIME FRAMEWORK

BayLIME uses Bayesian linear regressors as local surrogate models, and implements three options as shown in Table 1.

Table 1: The BayLIME framework with three options

<table>
<thead>
<tr>
<th>BayLIME Opt.</th>
<th>$\mu_0$</th>
<th>$\lambda$</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-informative priors</td>
<td>zero vector</td>
<td>fitted</td>
<td>fitted</td>
</tr>
<tr>
<td>partial informative priors</td>
<td>known</td>
<td>known</td>
<td>fitted</td>
</tr>
<tr>
<td>full informative priors</td>
<td>known</td>
<td>known</td>
<td>known</td>
</tr>
</tbody>
</table>

BayLIME with non-informative priors. When no prior knowledge is available, we assume a zero mean vector for $\mu_0$ and do Bayesian model selection [Bishop [2006]; Cht. 3.4] for $\lambda$ and $\alpha$. Specifically, BayLIME reuses the established algorithm to fit $\alpha$ and $\lambda$ from data [MacKay [1992]; Tipping [2001]. The solutions for $\lambda$ and $\alpha$ are implicit, since they are obtained by starting with initial values and then iterating over some interval equations until convergence.

BayLIME with partial informative priors. We assume a known complete prior distribution of the feature coefficients with mean vector $\mu_0$ and precision $\lambda$. We call this option in BayLIME “with partial priors” in the sense that we still do not know the parameter $\alpha$, for which we fit from data. Similarly as before, we modify the Bayesian model selection algorithm by iterating $\alpha$ (but with fixed $\lambda$ in this case) to maximise the log marginal likelihood until convergence.

BayLIME with full informative priors. By assuming an ideal scenario in which we have full prior knowledge of all the $\mu_0$, $\lambda$, and $\alpha$ parameters, BayLIME may directly implement the closed-form results of Eq. (8).

6 EVALUATION

We evaluate our BayLIME by performing extensive experiments to address the following research questions.

RQ1: How effectively BayLIME improves the consistency of LIME in different scenarios of available prior knowledge?
RQ2: How effectively BayLIME improves the robustness to kernel settings of LIME in different scenarios of available prior knowledge?
RQ3: Will BayLIME improve the explanation fidelity compared to some state-of-the-art XAI methods in different scenarios of available prior knowledge?

Evaluation methods. LIME is one of the few XAI techniques that work for all tabular data, text and images [Molnar [2020]. Our experiments were also conducted on diverse types of datasets including the Boston house-price dataset, the breast cancer Wisconsin dataset (with both numerical and categorical features), and a set of CNNs pretrained on the ImageNet and GTSRB [Stallkamp et al. [2012]. While

5.3 WAYS OF OBTAINING PRIOR KNOWLEDGE

How to elicit accurate priors for BayLIME varies case by case (depending on the application-specific context), and is indeed challenging. However, we believe such challenge is neither unique to BayLIME, but generic to any approach with a Bayesian flavour, nor clueless in practice. That said, we implemented 3 ways of getting priors in our experiments (but not limited to, cf. Sec. 8 for more discussions):

a) Validation and Verification (V&V) methods that directly analyse the behaviour of the underlying AI model may indicate the importance of certain features, yielding priors required by BayLIME. E.g., when explaining a prediction made by an infected model, (imperfect) detection tools may provide prior knowledge on possible backdoor triggers. We implemented this scenario by using the results of the popular NeuralCleanse [Wang et al. [2019] as the priors.

b) There are emerging XAI techniques (cf. Huang et al., 2020; Molnar [2020] for a review). Explanations by some other diverse XAI explainers based on fundamentally different theories to LIME (e.g., gradient-based vs perturbation-based, global vs local) may provide useful prior knowledge. Presumably, the drawbacks of individual XAI methods can be mitigated by the “hybrid” solution provided by BayLIME. We implemented and confirmed this conjecture by using GradCAM [Selvaraju et al. [2017] results as the priors.

c) We also use explanations of a set of similar instances (to the instance under explanation) to form our prior knowledge. Although it is still not rigorous enough (e.g., how to decide what is a “similar” instance), we believe this way of obtaining priors serves as a first illustrative example.
only typical results are presented in the main paper, cf. the appendices and our BayLIME project website for more.

For RQ1, in addition to LIME, we select a set of BayLIME explainers with different options and prior parameters. For each explainer, we iterate the explanation of the given instance \(k\) times, and quantify the inconsistency according to Kendall’s W. We treat the sample size \(n\) as a variable to confirm our earlier observation that inconsistency is an even severer problem when \(n\) has to be limited for efficiency.

To answer RQ2, we firstly define an interval \([l_0, l_p]\) as the empirical bounds of all possible kernel width settings for the given application. We randomly sample from that interval 5000 pairs of kernel width parameters. Then, for each pair, we can calculate the “distance” of the two explanations. Finally, we obtain a sample set of ratios between the two distances of explanations and the kernel width pair, on which statistics provides insights on the general robustness, e.g., the median value as used by Eq. (2).

The prior knowledge used in RQ1 and RQ2 is obtained from previous LIME explanations on a set of similar instances. Specifically, the average importance of each feature in that set collectively forms our prior mean vector \(\mu_0\), and the number of similar instances implies \(\lambda\), leaving the \(\alpha\) either as unknown (then fitted from data) or assigned empirically (based on previous fitting of \(\alpha\) on similar instances).

For RQ3, we implemented 2 scenarios in which the prior knowledge is obtained from other diverse XAI and V&V methods respectively:

**RQ3.a** We first use the gradient-based XAI approach GradCAM to obtain “grey-box” information of the underlying CNN as the prior, and then combines such extra knowledge with the “black-box” evidence as utilised by LIME. We compare the explanation fidelity of BayLIME with LIME, GradCAM and SHAP, via the deletion and insertion metrics.

**RQ3.b** In this scenario, we aim to explain the behaviour of infected CNNs (BadNet [Gu et al., 2019] and TrojanAttack [Liu et al., 2018] both trained on GTSRB with backdoors). The backdoor detector NeuralCleanse is used to approximate the location of triggers first. Then, BayLIME considers such knowledge as priors to provide a better explanation on the attacking behaviour, compared to NeuralCleanse and LIME applied solely. For the two example scenarios, we choose the relatively more practical option – BayLIME with partial priors, i.e., \(\alpha\) is fitted, and the readers are referred to Appendix A for the formulae with intuitive rationales on eliciting \(\mu_0\) and \(\lambda\).

Results and Discussion. RQ1. The red curves in Fig. 2 present the Kendall’s W measurements as a function of the perturbed sample size \(n\). Although it increases quickly, we observe very low consistency when \(n\) is relatively small (e.g., \(n < 200\)). These results support our earlier conjecture on the inconsistency issue of LIME, especially when \(n\) has to be limited by an upper-bound due to efficiency considerations. Non-informative BayLIME is indistinguishable from LIME, since both of them are only exploiting the information from the \(n\) samples that generated randomly – the presence of randomness means that the results of sampling cannot be duplicated if the process were repeated. Naturally, the more sparse the samples are, the greater randomness presents in the dataset. Thus, LIME and non-informative BayLIME show monotonic trends as \(n\) increases.

By contrast, BayLIME with partial/full informative priors can “average out” the sampling noise in the new generated data by combining the information from the priors, in a Bayesian principled way of Eq. (9). To closely inspect how effectively BayLIME with different priors affects the consistency, we first need the auxiliary of the factor \(\lambda/\alpha\). It is a known result that \(\lambda/\alpha\) can be treated as a regularization coefficient in Bayesian linear regressors [Bishop, 2006 Chpt. 3.3], meaning a larger \(\lambda/\alpha\) penalises more on the training data (to control over-fitting). Indeed, this aligns well with our Remark (i) when \(\alpha \approx 0\), the factor \(\lambda/\alpha\) → +∞ meaning a huge penalty on the data, thus the prior knowledge effectively dominates the posteriors; (ii) when \(\lambda \approx 0\), \(\lambda/\alpha \rightarrow 0\) meaning no penalty on the data, thus the posteriors is dominated by the new observations.

Both the plots of BayLIME with full informative priors (yellow curves) in Fig. 2 (A) and (C) have a regularization
factor $\lambda/\alpha = 20$, and are basically identical. This is because, once $\lambda/\alpha = 20$ is fixed, the “proportion” of contributions to the posteriors by the priors and the new data is fixed. In other words, given $n$ samples, the ability of “averaging out” sampling noise by the priors is fixed. When $\lambda/\alpha$ increases to 200, as shown by the yellow curves in Fig. 2(B) and (D), such ability of averaging out sampling noise is even stronger, which explains why Kendall’s W measurements in this case are higher than the case of $\lambda/\alpha = 20$. For BayLIME with partial informative priors (blue curves in Fig. 2), we observe that smaller $\lambda$ results in worse consistency, e.g., Fig. 2(C) vs (D). Again, Remark 1 applies here – smaller $\lambda$ implies less contributions from the priors to the posteriors, meaning with less ability to average out the randomness in the samples.

Starting from a non-zero small number, as $n$ increases, we can see BayLIME with partial/full informative priors may exhibit an uni-modal pattern, e.g., the Fig. 2(C) with a minimum point. This represents a tension between the perfectly consistent prior knowledge (does not change at all in repeated explanations) and quite consistent MLE based on large samples. There must be a “balance-point” in-between that compromises both ends of the tension, yielding a minimised consistency. Finally, when $n \to +\infty$, it is trivial to see (e.g., by taking the limit of Eq. (12) as a function of $n$), all plots will eventually converge (to the measurement based on MLE using infinitely large samples).

RQ2. Fig. 3 are box-and-whisker charts providing insights on the general robustness of eight AI explainers to kernel width settings, in which the median values defined by Eq. (2) are marked, as usual, by horizontal bars inside the boxes. Again, LIME and BayLIME with non-informative priors exhibit similar robustness, since there is no prior knowledge being considered, rather the data solely determines the explanations of both. In stark contrast, when either partial or full prior knowledge is taken into account, we observe an obvious improvement on the robustness to kernel settings.

![Figure 3: The general (un-)robustness of eight AI explainers to kernel settings (box-and-whisker plots without outliers).](image)

The regularisation factor $\lambda/\alpha$ and Remark 1 are still handy here in the discussions on how varying the $\lambda$ and $\alpha$ affects the robustness – it all boils down to how much contribution from the priors (that is independent from kernel setting), compared with the contribution from the new data (that is sensitive to kernel settings), to the posteriors.

RQ3.a. GradCAM provides fundamentally different grey-box information as priors to the black-box LIME. BayLIME, as expected, performs better than both GradCAM (the prior) and LIME (new observations), thanks to its unique advantage of incorporating diverse information. Fig. 4 first shows two sets of such examples via the AUC scores of the deletion (smaller is better) and insertion (bigger is better) metrics of BayLIME comparing with other XAI methods.

![Figure 4: Two sets of examples ($n = 300$), and the average AUC (based on 1000 images per value of $n$) of the deletion (smaller is better) and insertion (bigger is better) metrics of BayLIME comparing with other XAI methods.](image)

Again, LIME and BayLIME with non-informative priors exhibit similar robustness, since there is no prior knowledge being considered, rather the data solely determines the explanations of both. In stark contrast, when either partial or full prior knowledge is taken into account, we observe an obvious improvement on the robustness to kernel settings.

RQ3.b. The V&V tool NeuralCleanse yields reversed trig-
Figure 5: Two examples and their IoU measurements of BayLIME using V&V results as priors (reversed triggers by NeuralCleanse), compared to the results of NeuralCleanse and LIME applied individually. LIME and BayLIME use the top-few features (where each feature is a fixed size square superpixel, cf. Appendix A.2 for details) such that the total number of pixels is the same as the ground truth. NB: We omit the background image to highlight the triggers only.

Table 2: Statistics on the average IoU by 3 methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>NeuralCleanse</th>
<th>LIME</th>
<th>BayLIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>BadNet</td>
<td>0.000</td>
<td>0.385</td>
<td>0.406</td>
</tr>
<tr>
<td>TrojanAttack</td>
<td>0.150</td>
<td>0.599</td>
<td>0.637</td>
</tr>
</tbody>
</table>

8 DISCUSSION

Prior knowledge plays a key role in BayLIME, thus we discuss the following questions on priors to highlight its practical usefulness.

Where can we obtain the priors? The answer to this question varies case by case, depending on the application-specific context. We have implemented three typical ways of obtaining priors, i.e. from other diverse XAI tools, V&V methods, and previous explanations of similar instances. More practical ways of getting priors for BayLIME are certainly possible and we plan to investigate more in future.

How do we know if the prior is good? First of all, we need to be clear about the definition of “good” priors – is it the prior that truly/accurately reflects the unknown behaviour of the AI model? Or the prior gives an explanation that looks good to human users? The latter definition is “cheating” in a Bayesian sense – we cannot rig a prior to get a result we like – and this is not the purpose of XAI methods. The former is sensible, but due to the lack of ground truth behaviours of the black-box AI model, we would never know the prior is good or not with certainty. However, we can be fairly confident that a given prior is good in some practical cases (e.g., derived from a V&V tool that was shown to be reliable in previous uses), so that it can be utilised by BayLIME.

What if we used a bad prior? Reusing the sensible definition of good priors above, indeed, there could be the case we are using a bad prior that introduces bias. In this case,
BayLIME might end up providing an explanation that is “consistently and robustly bad”, which seems to imply that BayLIME is only useful when we are certain that the priors are good. However, in practice, we can never be certain if a prior is good or bad, rather the prior is simply a piece of evidence to us. It is not only against the Bayesian spirit but also unwise to discard any evidence without sound reasons. If there is a proof that the evidence (either the priors or the new observations) is not trustworthy, then certainly we should not consider it in our Bayesian inference – in this sense, all Bayesian methods depend on good priors, not just BayLIME. So with some caution in deriving priors/evidence from trustworthy ways in practice (a separate engineering problem that is normally out of the scope of a Bayesian method itself), we may ease the concern of BayLIME being “consistently and robustly bad”.

9 CONCLUSIONS AND FUTURE WORK

This paper introduces BayLIME, a principled, Bayesian based method to integrate useful prior knowledge into the well-known XAI method LIME. Our theoretical analysis and extensive experiments show that the Bayesian mechanism helps in improving the consistency in repeated explanations of a single prediction, robustness to kernel settings, and the explanation fidelity.

Embedding diverse sources of knowledge has been a clear trend in AI in recent years, as the effort to make black-box learning methods more transparent and interpretable. Emerging techniques peeking inside AI models will lead to many different ways of obtaining knowledge of their black-box behaviour. Therefore, principled methods to integrate diverse knowledge are called for.

In future work, we will investigate more practical ways of obtaining priors and how to leverage the posterior confidence for better explanations. Last but not least, we plan to develop novel use cases of BayLIME.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 956123. This work is supported by the UK EPSRC (through the Offshore Robotics for Certification of Assets [EP/R026173/1] and End-to-End Conceptual Guarding of Neural Architectures [EP/T026995/1]) and the UK Dstl (through the project Safety Argument for Learning-enabled Autonomous Underwater Vehicles). XZ’s contribution is partially supported through Fellowships at the Assuring Autonomy International Programme.

References


