An Additive Instance-Wise Approach to Multi-class Model Interpretation

Vy Vo[1] **Van Nguyen**[2] **Trung Le**[1]
Quan Hung Tran[3] **Gholamreza Haffari**[1] **Seyit Camtepe**[4] **Dinh Phung**[1]

[1]Monash University, Australia
[2]Adelaide University, Australia
[3]Adobe Research, San Jose, CA, USA
[4]CSIRO Data61, Australia

tvoo0019@student.monash.edu, khacvan.nguyen@adelaide.edu.au,
{trunglm,gholamreza.haffari,dinh.phung}@monash.edu,
qtran@adobe.com, seyit.camtepe@data61.csiro.au

Abstract

Interpretable machine learning offers insights into what factors drive a certain prediction of a black-box system and whether to trust it for high-stakes decisions or large-scale deployment. Existing methods mainly focus on selecting explanatory input features, which follow either locally additive or instance-wise approaches. Additive models use heuristically sampled perturbations to learn instance-specific explainers sequentially. The process is thus inefficient and susceptible to poorly-conditioned samples. Meanwhile, instance-wise techniques directly learn local sampling distributions and can leverage global information from other inputs. However, they can only interpret single-class predictions and suffer from inconsistency across different settings, due to a strict reliance on a pre-defined number of features selected. This work exploits the strengths of both methods and proposes a global framework for learning local explanations simultaneously for multiple target classes. We also propose an adaptive inference strategy to determine the optimal number of features for a specific instance. Our model explainer significantly outperforms additive and instance-wise counterparts on faithfulness while achieves high level of brevity on various data sets and black-box model architectures.

1 Introduction

Black-box machine learning systems, such as kernel methods or deep neural networks, enjoy remarkable predictive performance at the cost of interpretability. This trade-off has motivated a number of interpreting approaches for explaining the reasoning process underlying these complex models. Such explanations are particularly useful for healthcare (Caruana et al., 2015; Rich, 2016), cybersecurityNguyen et al. (2021) or criminal investigation (Lipton, 2018), where intelligible prediction and transparent decision-making are crucial. Though model interpretation can be done through various approaches: feature importance, visualization, or examples (Mothilal et al., 2020), our work focuses on the first one, that is to assign relative importance weights to individual features with respect to the model's prediction on an input instance. In the context of model explanation, features refer to input components interpretable to humans. For high-dimensional data such as texts or images, features can be a bag of words/phrases or a group of pixels/super-pixels (Ribeiro et al., 2016). Explanations are generally made by selecting the top K features with the highest weights, signifying K most influential features to a black-box's decision. There are two main directions to learning feature importance: **Locally Additive** (Simonyan et al., 2014; Bach et al., 2015; Springenberg et al., 2014; Shrikumar et al., 2017; Baehrens et al., 2010; Ribeiro et al., 2016; Lundberg and Lee, 2017) and **Instance-Wise**. (Chen et al., 2018; Bang et al., 2019). The most popular *additive* method is LIME

(Ribeiro et al., 2016). LIME explains the prediction for an input instance by randomly sampling perturbations around it and using them to locally approximate the black-box decision boundary. LIME thus suffers from high variance due to perturbation randomness (Slack et al., 2020; Situ et al., 2021), and there is chance that the perturbed examples behave undesirably, for example, to change the prediction of the original model. Additive models are also highly inefficient because they learn individual explainers for every instance. The second line of works approaches the problem from an information-theoretic perspective. These methods, including L2X (Chen et al., 2018) and VIBI (Bang et al., 2019), produce a feature selector that returns a local distribution over input features. The feature selector is trained globally by maximizing the mutual information between the selected features and response variable. After training, explanations can be obtained simultaneously and efficiently for multiple instances. The technique is thus referred to as *instance-wise* feature selection. There is however one drawback about this setup: K is treated as a hyperparameter and careful tuning for K is essential. Different choices of K can yield different feature rankings e.g., the features picked by a model trained on top 5 may be considered irrelevant by a model trained on top 10 (See Appendix H.2).

Contributions. This work attempts to integrate both approaches into an **additive instance-wise** framework that simultaneously tackles all issues discussed above. Our explainer is not explicitly trained to select the top K features, rather to learn the local attributions of features across the multi-class output space. The attributions are captured in a spatial module denoted as W, which interacts with input instances in a locally additive manner. This module is globally optimized together with a learnable feature selector using the information-based approach. The motivation is to have local distributions learned to generate high-quality local samples, thereby addressing LIME's problem related to heuristic sampling. In Section 5, we show that an explainer that learns local distributions performs significantly better than one trained on heuristic examples. Importantly, our setup does not require tuning for K as done in the existing works. K in fact can be determined during the inference stage and particularly adaptive to each sample.

Our contributions are summarized as follows

- We introduce **AIM** - an **A**dditive **I**nstance-wise approach to **M**ulti-class model interpretation. Our model explainer inherits merits from both family of methods: model-agnosticism, flexibility while supporting efficient interpretation for multiple decision classes.
- We propose an adaptive inference strategy that not only enables the selection of K most relevant features but also outputs the optimal value of K specifically to each input example.
- Our model explainer is shown to produce remarkably faithful explanations of high quality and compactness. Through quantitative and human assessment results, we achieve superior performance over the baselines on different data sets and architectures of black-box model.

2 Related work

Early additive methods employ gradient values to estimate importance scores. Without any training, the process involves back-propagation for calculating the gradients of the output neuron with respect to the input features. Saliency Maps (Simonyan et al., 2014) are constructed accordingly, highlighting object locations relevant to a specific target class. This approach however suffers from vanishing gradients during the backward pass through ReLU layers, which may downgrade important features. Several methods are proposed to improve the propagation rule (Bach et al., 2015; Springenberg et al., 2014; Shrikumar et al., 2017). While the aforementioned papers solely focus on convolution neural networks, Baehrens et al. (2010) targets a broader family of classifiers using class probability gradients. When the black-box model is unknown or too complex, Parzen windows approximation is applied. Other model-agnostic methods make use of perturbations, notably LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). Given an input instance, perturbed samples are uniformly drawn within its neighborhood. Whereas LIME fits these examples to a linear separator and estimates feature weights through the coefficients, SHAP measures feature importance via the Shapey values computed from the model's conditional expectation function. To avoid local additivity that operates on instance-specific level, later works (Chen et al., 2018; Bang et al., 2019) utilize information theory in an instance-wise framework. In this approach, explanations are made by selecting features according to a logit vector. It is obtained from a globally trained feature selector with each logit score representing the attribution of a feature. Recently, there is a shift in interest towards generating explanations with certain properties, such as stability (Situ et al., 2021), sufficiency (Carter et al., 2019) or counterfactuals (Schwab and Karlen, 2019; Mahajan et al., 2019; Mothilal et al., 2020).

3 Method

3.1 Problem Setup

In the scope of this paper, we limit the current discussion to classification problems. However, the framework can also be adapted for regression problems. Consider a data set of pairs (X, Y) where $X \sim \mathbb{P}_X(\cdot)$ is the input random variable and Y is characterized by the conditional distribution $\mathbb{P}_m(Y \mid X)$ obtained as the predictions of a pre-trained black-box model for the response variable. We denote $x \in \mathbb{R}^d$ as an instance realization with d interpretable features and predicted label $Y = c \in \{1, ..., C\}$. Given the instance x, we can obtain the hard prediction from the black-box model as $y_m = \mathrm{argmax_c}\ \mathbb{P}_m(Y = c \mid X = x)$.

Earlier methods generate a single $d-$dimensional weight vector w_x assigning the importance scores to each feature of x. By training individual explainers, LIME (Ribeiro et al., 2016) in particular only provides a local explanation and cannot take advantages of the global information in the training set. To enable global behavior effectively (e.g., two similar instances should have similar explanations), our model aims to learn the attributions of features across all data instances in a training set. Specifically, we define an **explainer** $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}^{d\times C}$ mapping an input x to a weight matrix $W_x \in \mathbb{R}^{d\times C}$ with the entry $W_x^{i,j}$ representing the relative weights of the ith feature of x to the predicted label $j \in \{1, ..., C\}$.

$$\mathcal{E}(x) = W_x$$

Let $z \in \{0, 1\}^d$ be a random variable with the entry $z^i = 1$ indicating the feature ith is important to the black-box's predictions. With respect to x, we employ a **selector** $\mathcal{S} : \mathbb{R}^d \rightarrow [0, 1]^d$ that outputs $\mathcal{S}(x) = \pi_x$ such that $\pi_x^i := \mathbb{P}(z^i = 1 \mid X = x), i = 1, ..., d$. In other words, we can imagine π_x^i as the probability that the ith feature of x appears in the set of selected attributions.

3.2 Training Objectives

Intuitively, if the feature ith of x contributes more to the predictions of the black-box model, i.e., $\pi_x^i \approx 1$, the explainer is expected to give higher assignments to the row vector $W_x^{i,:}$. To mimic how the black-box behaves towards different attributions, we minimize cross entropy $H\big(\mathbb{P}_{\mathcal{E}}, \mathbb{P}_m\big)$ over quantities "weighted" by π_x, giving rise to the following objective

$$\mathcal{L}_1 = \mathbb{E}_x\Big[\mathrm{CE}\left(\bar{y}_m, \sigma(W_x^T\pi_x)\right)\Big] \tag{1}$$

where CE is the cross-entropy loss, $\bar{y}_m = \mathrm{argmax}_c\ \mathbb{P}_m(Y = c \mid x \odot \pi_x)$ with the element-wise product $\odot$, and σ is the softmax operator. Here note that we use $\sigma(W^T\pi_x)$ to define $\mathbb{P}_{\mathcal{E}}(Y \mid X = x)$ and minimize $H\big(\mathbb{P}_{\mathcal{E}}, \mathbb{P}_m\big)$ to approximate the black-box model distribution $\mathbb{P}_m$ in a hope of selecting good attributions for x. This generally concurs with LIME. However, different from LIME, we feed the probability vector π_x directly to the linear classifier with the weight matrix W_x rather than a binary vector $z_x \sim \mathrm{MultiBernoulli}(\pi_x)$ (i.e., $z_x^i \sim \mathrm{Bernoulli}(\pi_x^i), i = 1, ..., d$). The advantage is twofold. First, we can use the loss $\mathcal{L}_1$ to update directly the explainer $\mathcal{E}$, avoiding using Gumbel-Softmax trick (Jang et al., 2016; Maddison et al., 2016) to relax z_x. Second, we empirically find that this strategy works better and more stably than feeding relaxed samples (See Appendix C).

Jointly training $\mathcal{E}$ and $\mathcal{S}$ on the above criterion alone is difficult to yield effective probabilities π_x. We find that the selector tends to get "lazy" by producing $\pi_x \approx \mathbf{1}$ (i.e., the vector of all 1's), while we expect π_x is a sparse vector, wherein the elements corresponding to the important attributions are close to 1. To solve this, we adopt the information-theoretic perspective. Let x_S denote the sub-vector formed by the subset of K most important features $S = \{i_1, \dots, i_K\} \subset \{1, \dots, d\}$ $(i_i < i_2 < \dots < i_K)$. Given a random vector $X_S \in \mathbb{R}^K$, we maximize the mutual information.

$$\mathbb{I}(X_S; Y) = \mathbb{E}\Big[\log\frac{\mathbb{P}_m(Y \mid X_S)}{\mathbb{P}_m(Y)}\Big] = \mathbb{E}_X\mathbb{E}_{S|X}\mathbb{E}_{Y|X_S}\Big[\log\mathbb{P}_m(Y \mid X_S)\Big] + \text{Constant} \tag{2}$$

Our motivation concurs with L2X (Chen et al., 2018), wherein by maximizing $\mathbb{I}(X_S; Y)$, we encourage the selector $\mathcal{S}$ to learn and produce a meaningful π_x, focusing more on good attributions for x. Using the following inequality, we can obtain a variational lower bound for $\mathbb{I}(X_S; Y)$ by a generic choice of conditional distribution $\mathbb{Q}_S(Y \mid X_S)$

$$\begin{aligned}\mathbb{E}_{Y|X_S}\left[\log \mathbb{P}_m\left(Y \mid X_S\right)\right] &= \mathbb{E}_{Y|X_S}\left[\log \mathbb{Q}_S\left(Y \mid X_S\right)\right] + \mathrm{KL}\left(\mathbb{P}_m\left(Y \mid X_S\right), \mathbb{Q}_S\left(Y \mid X_S\right)\right) \\ &\geq \mathbb{E}_{Y|X_S}\left[\log \mathbb{Q}_S\left(Y \mid X_S\right)\right]\end{aligned}$$

where KL represents the Kullback-Leibler divergence.

Maximizing the mutual information in (2) can be therefore relaxed to maximizing the variational lower bound $\mathbb{E}_X \mathbb{E}_{S|X} \mathbb{E}_{Y|X_S}\Big[\log \mathbb{Q}_S(Y \mid X_S)\Big]$. We parametrize $\mathbb{Q}$ with a neural network approximator $\mathcal{G}$ such that $\mathbb{Q}_S(Y \mid x_S) := \mathcal{G}(x_S)$. We approximate x_S with an element-wise product $z_x \odot x$. More concretely, if x contains discrete features (e.g., words), we embed a feature (e.g., a selected word) in S with an learnable embedding vector, wherein a feature not in S is replaced with a zero vector.

To make the process continuous and differential for training, the Gumbel-Softmax trick (Jang et al., 2016; Maddison et al., 2016) is applied for relaxing Bernoulli variables z_x^i. In particular, the continuous representation $\tilde{z}_x^i$ is sampled from the Concrete distribution as $\left[\tilde{z}_x^i, 1 - \tilde{z}_x^i\right] \sim$ Concrete$(\pi_x^i, 1 - \pi_x^i)$:

$$\tilde{z}_x^i = \frac{\exp\{\big(\log \pi_x^i + G_{i1}\big)/\tau\}}{\exp\{(\log(1 - \pi_x^i) + G_{i0})/\tau\} + \exp\{(\log \pi_x^i + G_{i1})/\tau)\}}$$

with temperature τ, random noises G_{i0} and G_{i1} independently drawn from **Gumbel** distribution $G_t = -\log(-\log u_t),\ u_t \sim$ **Uniform**$(0, 1)$. The reason why we use continuous representations $\tilde{z}_x$ instead of π_x as in $\mathcal{L}_1$ is that $\mathcal{G}$ can be so expressive that $\mathbb{Q}_S$ becomes very close to $\mathbb{P}_m$ and would not support learning sparse π_x.

The optimization problem in (2) is transformed into minimizing the criterion below,

$$\mathcal{L}_2 = \mathbb{E}_x\Big[\mathrm{CE}\left(y_m, \mathcal{G}(\tilde{z}_x \odot x)\right)\Big] \tag{3}$$

The final objective. All of the networks $\mathcal{E}, \mathcal{S}$ and $\mathcal{G}$ are jointly learned over total parameters θ. We further introduce a regularization term over W to encourage sparsity, thereby assuring compact explanations. The final objective function is now given as,

$$\min_\theta \Big[\mathcal{L}_1 + \alpha\, \mathcal{L}_2 + \beta\, ||W||_{2,1}\Big]$$

where $\|\cdot\|_{2,1}$ is the group norm $2, 1$, and α, β are balancing coefficients on loss terms. Additionally, α and β are subject to tuning since a highly compressed representation can cause information loss and harm faithfulness.

3.3 Inference

In the existing works, an explainer produces, for an input instance x, a single vector assigning importance scores to each feature. The **conventional inference** method is to choose top K features with the highest weights, with K determined in advance for both training and inference of explanations. In our framework, the explainer outputs a weight matrix W_x size $d \times C$ (recall that d is the number of features and C is the number of target classes). We obtain the black-box's predicted label $j = y_m = \mathrm{argmax_c}\, \mathbb{P}_m(Y = c \mid X = x)$ and select the corresponding column $W_x^{:,j}$ as the weight vector. Features can then be derived accordingly. However, we argue that for most applications, end-users are interested in instance-specific explanations, and the instance may require a larger or smaller number of features, than a fixed set K, to approximate the black-box well. We therefore propose an **adaptive inference** strategy to decide the optimal K for each individual input instance.

Specifically, we first set a pre-defined budget $K_{max} \leq d$ and sort the weight vector in the descending order. For each K from 1 to K_{max}, we create a masked input variant by concatenating the sub-vector

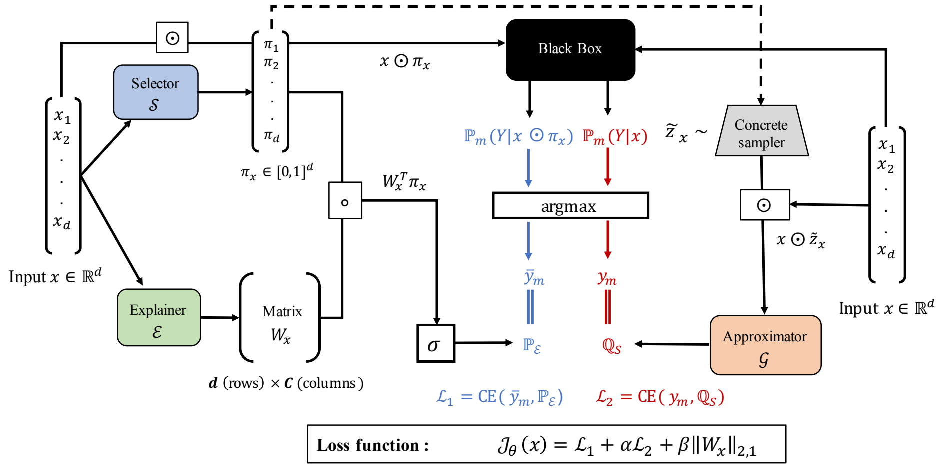

Figure 1: Illustration of AIM pipeline. CE is cross-entropy loss function and σ is Softmax operator. Our framework consists of an explainer $\mathcal{E}$, a feature selector $\mathcal{S}$, and an approximator $\mathcal{G}$. The objective $\mathcal{L}_1$ encourages the explainer to mimic the black-box by aligning their predictions on the input wherein each feature is "weighted" by its importance score. The objective $\mathcal{L}_2$ supports the learning of local distributions π_x and ensures the selected features reserve the most information of the original input.

of K-largest-weighted features with the zero vector of size $d-K$ corresponding to $d-K$ unselected elements. We then feed this masked input variant to the black-box model and apply an early stopping when the predicted label for the masked input variant is exactly the original hard prediction y_m. The early-stopping K can be regarded as the optimal K for an input instance. This will be done sequentially for all examples in a dataset, yielding two kinds of explanation: the optimal number of features is required to explain the example and the subset of important features itself. When d is large, one challenge is to how to set K_{max} to be reasonably low in order to ensure efficiency yet not to sacrifice faithfulness. Fortunately, our experiments show that, under adaptive inference, an explainer can achieve faithfulness of over 90% at very small values of K_{max}.

4 Experiments

We conduct experiments on 3 text classification and 2 image recognition tasks. Here we mainly discuss the setup and results for the text classifiers. Details about our experiments on MNIST (LeCun et al., 1998) and Fashion-MNIST datasets (Xiao et al., 2017) are provided in Appendix F.

- **Sentiment Analysis:** The Large Movie Review Dataset **IMDB** (Maas et al., 2011) consists of 50,000 movie reviews with positive and negative sentiments. The black-box classifier is a bidirectional GRU (Chen et al., 2018) that achieves a 85.4% test accuracy.
- **Hate Speech Detection: HateXplain** is an annotated dataset of Twitter and Gab posts for hate speech detection (Mathew et al., 2021). The task is to classify a post either to be normal or to contain hate/offensive speech. The black-box model is a bidirectional LSTM (Gers et al., 2000) stacked under a standard Transformer encoder layer (Vaswani et al., 2017) of 4 attention heads. The best test accuracy obtained is 69.6%.
- **Topic Classification:** AG is a collection of more than 1 million news articles. **AG News** corpus (Zhang et al., 2015) is constructed by selecting 4 largest classes from the original dataset: World, Sports, Business and Sci/Tech. We train a word-level convolution neural network (CNN) (LeCun et al., 1995) as a black-box model. It obtains 89.7% accuracy on the test set.

Dataset statistics are provided Appendix A. Code and data for reproducing our experiments are anonymously published at `https://github.com/isVy08/AIM`.

4.1 Performance Metrics

An essential criterion of a model explainer is faithfulness, generally referring to how well the explainer mimics the black-box model it interprets. Despite the lack of a standard definition, we follow the Model Assumption, stated in Corollary 1.2 of Jacovi and Goldberg (2020), defining an explanation to be faithful if it produces the same decision from the original model. In addition, we study the quality of explanations by examining brevity and purity of the selected features. The metrics for quantitative evaluation are presented as follows:

- **Faithfulness:** Given the weight vector produced by the explainer, for a testing sample x, we select a subset of most important features $\mathcal{S}$. As in Section 3, we approximate $x_{\mathcal{S}}$ with a vector $\tilde{x}$ where unselected features are masked by zero paddings. We fed $\tilde{x}$ to the black-box model and compute the accuracy against the black-box model's original predictions on the full document. Higher accuracy indicates the features selected are representative of the input and strongly relevant to the model's prediction. [1]
- **Brevity:** Given a subset $\mathcal{S}$, we define an explainer achieves brevity if the subset contains closely related features. For textual data, we expect the chosen features contain a large number of duplicates and/or synonyms. We introduce *cluster ratio* to quantify brevity. Specifically, we first collect a database of semantically related words through WordNet (Miller, 1995). We group tokens in $\mathcal{S}$ into clusters of synonyms, then calculate the average number of clusters formed over K tokens. Larger ratio means the tokens are more semantically polarizing and diverse. We thus want this ratio to be small.
- **Purity:** For text classification tasks, we observe that an explainer sometimes selects stopwords or punctuation as important features, which are meaningless from an end-user's perspective. An effective explainer should reduce the likelihood of picking such "contaminated" features. We therefore investigate *the proportion of stopwords and punctuation* included in $\mathcal{S}$. High proportion is equivalent to a low-quality feature set.

4.2 Baseline Methods

We compare our model against three baselines: L2X (Chen et al., 2018), LIME (Ribeiro et al., 2016) and VIBI (Bang et al., 2019). The experiments for the baselines are run on the authors' published codes. All the baselines are trained at $K = 10$ and we use the default architectures reported by the authors for all datasets. VIBI further offers multiple options for the approximator. In our experiments, LSTM approximator gives the highest accuracies for both IMDB and AG News, while CNN works best on HateXplain. For training LIME, we have 2000 perturbations generated around per instance, and though LIME can explain a single target class, we use the multi-class setup to keep it consistent with the other baselines. It is worth noting that our black-box architecture for IMDB dataset is different from ones reported in Chen et al. (2018) and Bang et al. (2019). In their experiments, Chen et al. (2018) adopts a CNN (Zhang et al., 2015) while Bang et al. (2019) chooses a hierarchical LSTM. These architectures are similar to those of their own explainers, which we suspect might have some influence on the explainers' performance. Since our explainer is not GRU-based either, we intentionally opt for a bidirectional GRU in order to examine whether these models can explain different kinds of black-box architectures. Apart from LIME that runs on a single CPU, the other models including AIM are trained on 4 NVIDIA Tesla V100 GPUs.

4.3 Model Design

We parametrize $\mathcal{E}, \mathcal{S}$ and $\mathcal{G}$ by three deep neural network functions. Since our input X is discrete, every network contains a learnable embedding layer. The explainer $\mathcal{E}$ passes the embedded inputs into three 250-dimensional dense layers and outputs W after applying ReLU non-linearity. The selector $\mathcal{S}$ is composed of one bidirectional LSTM of 100 dimension and three dense layers of the same size. Each layer is stacked between a Dropout layer (Srivastava et al., 2014) and an activation. The upper layers use ReLU while Sigmoid is a natural choice for the final one. Regarding the network $\mathcal{G}$, after feeding the inputs into its own embedding layer, we process the outputs through a 250-dimensional

[1] In L2X, Chen et al. (2018) refers to this metric *post-hoc accuracy*. Bang et al. (2019) measures faithfulness differently for VIBI. They compare the predictive performance of the approximator with the black-box while we conduct a post-hoc evaluation of the black-box's performances given the original and masked inputs.

convolutional layer with kernel size 3, followed by a max-pooling layer over the sequence length. The last layer is simply a dense layer of dimension 250 together with Softmax activation. We use the same architecture for all tasks and train our model with Adam optimizer (Kingma and Ba, 2014) at $\tau = 0.2$ and learning rate of 0.001. We tune the coefficients α, β via grid search to achieve an adequate balance of faithfulness and compression. The best hyperparameters are reported in Appendix A.

4.4 Results

We train each model explainer with 5 different parameter initializations and report the average results on the test sets in Table 1. The evaluation is conducted under both conventional inference and adaptive inference as discussed in Section 3.3. We further report the mean value of K obtained when adaptive inference is implemented over all testing examples. Regarding the baselines, we apply the conventional method as originally conducted by the authors. Since the competing models are trained at $K = 10$, we set $K_{max} = 10$ during adaptive inference and validate the top 10 features for conventional inference.

Table 1: Performance of model explainers for 3 datasets. a and c denote model versions respectively under adaptive and conventional inference methods. ↑ Higher is better. ↓ Lower is better. As observed in examples of Table 2, **AIM**^{a} finds $1 - 2$ words as the explanation across the datasets, usually adjectives like *great*, *worst* etc.

Explainer	**AIM**^{a}	**AIM**^{c}	**L2X**^{c}	**LIME**	**VIBI**^{c}
			IMDB		
Mean K	1.15	10	10	10	10
Purity ↓	$\mathbf{3.23\% \pm 0.28\%}$	$9.66\% \pm 0.72\%$	$15.24\% \pm 1.04\%$	$40.62\% \pm 0.06\%$	$29.95\% \pm 0.08\%$
Brevity ↓	$\mathbf{1.03 \pm 0.00}$	2.06 ± 0.03	2.84 ± 0.04	4.54 ± 0.01	3.83 ± 0.02
Faithfulness ↑	$\mathbf{99.98\% \pm 0.02\%}$	$99.98\% \pm 0.02\%$	$85.97\% \pm 0.25\%$	$87.90\% \pm 0.17\%$	$56.20\% \pm 0.30\%$
			HateXplain		
Mean K	1.85	10	10	10	10
Purity ↓	$\mathbf{6.76\% \pm 0.86\%}$	$25.27\% \pm 1.60\%$	$22.89\% \pm 0.65\%$	$37.72\% \pm 0.05\%$	$26.43\% \pm 1.69\%$
Brevity ↓	$\mathbf{1.34 \pm 0.05}$	4.16 ± 0.10	4.41 ± 0.06	4.71 ± 0.01	3.75 ± 0.17
Faithfulness ↑	$\mathbf{98.86\% \pm 0.32\%}$	$96.06\% \pm 0.91\%$	$74.05\% \pm 0.40\%$	$85.47\% \pm 0.15\%$	$60.60\% \pm 1.66\%$
			AG News		
Mean K	1.56	10	10	10	10
Purity ↓	$\mathbf{0.64\% \pm 0.10\%}$	$4.47\% \pm 0.33\%$	$5.87\% \pm 0.45\%$	$21.52\% \pm 0.14\%$	$18.87\% \pm 0.51\%$
Brevity ↓	$\mathbf{1.14 \pm 0.01}$	3.35 ± 0.01	4.03 ± 0.05	4.75 ± 0.00	3.84 ± 0.05
Faithfulness ↑	$\mathbf{99.48\% \pm 0.10\%}$	$97.54\% \pm 0.19\%$	$88.75\% \pm 0.24\%$	$85.83\% \pm 0.49\%$	$62.98\% \pm 1.01\%$

Under both inference techniques, our model explainer remarkably achieves over 96% of faithfulness across all datasets with the highest level of brevity. The adaptive versions significantly surpass the baselines and the conventional counterparts on all metrics. Appendix H.1 provides an additional study on the performance of L2X under adaptive inference. Though the strategy works equally better for L2X than under conventional inference, AIM consistently outperforms L2X since our explainer is not dependent on K, nor does it require tuning K as a hyper-parameter. While conventional inference is naturally fast, one may be concerned about the efficiency of adaptive inference, especially for bulky black-box models. Our reported values for mean K reveals that the one most important feature is sufficient to explain an example on average, implying that only 1 iteration is required to converge.

Figure 2 illustrates the entire distributions of K optimal features upon adaptive inference. 98% of the examples require fewer than 8 features to approximate the black-box models, highlighting that adaptive inference is not computationally intensive. Appendix B further analyzes the performance vs. efficiency trade-off of our adaptive inference strategy in comparison with LIME. For AIM, we examine the CPU times vs. levels of faithfulness at various values of K_{max}, whereas for LIME, we are interested in how the these factors vary when increasing the number of training perturbations. At higher values

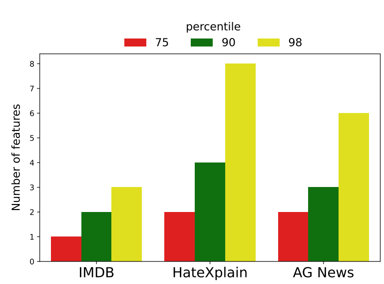

Figure 2: The full distributions of the values of optimal K when running AIM under adaptive inference without stopping at K_{max} (i.e., $K_{max} = d$).

of K_{max}, AIM becomes more faithful while computing times remain relatively stable. This implies that setting K_{max} at lower values for efficiency does not compromise model performance, and even at $K_{max} = 1$, faithfulness already goes above 97%. On the other hand, in order to obtain higher faithfulness, LIME must fit the model to a larger number of perturbations at the cost of an exponential increase in computing time. To achieve around 90%, LIME requires as many as 4000 samples per instance with the total processing time of 2^{14} seconds on a single CPU.

Table 2 provides 4 examples of the features chosen by AIM in IMDB dataset. These features strongly align with the predictions of the black-box model. In this regard, our explainer can help shed light on why the black-box makes such predictions, especially the wrong ones. We also find that leveraging global information helps us achieve more stable explanations than the baselines, with **stability** defined as the ability to produce the same explanations given similar examples (Situ et al., 2021). Appendix D provides evidence for this claim. A comprehensive qualitative comparison with the baselines on multiple examples can further be found in Appendix E. [2] While explanations from LIME are contaminated with a larger volume of neutral words, instance-wise methods tend to select more meaningful features. It is also clear AIM assures compactness by picking up all duplicates and synonyms without compromising predictive performance. This behavior is also observed in L2X and VIBI; however, these explainers are shown to be less faithful and perform inconsistently across datasets. Considering their architectures, the goodness of features heavily depends on the predictive performance of the explainer. If the explainer predicts a different label from the black-box, it would yield a wrong explanation, thus losing faithfulness. We therefore avoid this issue by directly choosing the feature attributions corresponding to the class predicted by the black-box model. Section 5 analyzes the scenario where the features are inferred according to the explainer.

Table 2: Ground-truth labels and labels predicted by the black-box model are given in the first two columns. 10 most relevant words selected by AIM are highlighted in yellow, with the optimal adaptive features bold in red.

Truth	Model	Key words
positive	positive	this movie was a pleasant surprise for me. in all honesty, the previews looked horrible, up until the point where emma thompson and alan rickman appeared. so i rented it with reservation, but i thoroughly enjoyed this movie. it had **great** acting, a few good plot twists, and, of course, emma thompson and alan rickman. it's definitely worth checking out.
negative	negative	this may just be the **worst** movie ever produced. **worst** plot, worst acting, worst special effects...be prepared if you want to watch this. the only way to get enjoyment out of it is to light a match and burn the tape of it, knowing it will never fall into the hands of any sane person again.
positive	negative	to me, "anatomie" is certainly one of the better movies i have seen. i don't think "anatomie" was primarily intended to be a horror movie but a movie questioning the ethics of science. if you watch it with that in mind, it turns into a really good film. the only annoying bit was the **awful** voice dubbing for the english version. how can you expect any non-german person to listen to these unbearable german accents for two hours ? let native english speakers do the talking or use subtitles instead!!
negative	positive	i have seen this movie several times, it sure is one of the cheapest action flicks of the eighties. so, i think many viewers would definitely change the channel when they come across this one. but, if you are into **great** trash, "dragon hunt" is made for you. the main characters (the mcnamara twins) are sporting great moustaches and look so ridiculous in their camouflage dresses. one of the best scenes is when one of then gets shot in the leg and is still kicking his enemies into nirvana. this movie is really awful, but then again, it is a great party tape!

4.5 Human Evaluation

We additionally conduct a human experiment to evaluate whether the words selected as an explanation convey sufficient information about the original document to human users. We ask 30 university students to infer the sentiments of 50 IMDB movie reviews, given only 10 key words obtained from an explainer for each review (outputs of conventional inference). To avoid confusion, only examples where the black-box model predicts correctly are considered (See Appendix G for the setup and the survey interface).

We assess whether the sentiment inferred by human is consistent with the actual label of a movie review: *human accuracy*. Some reviews

Table 3: Human evaluation results on IMDB dataset of the AIM, L2X, and LIME methods.

Explainer	**AIM**	**L2X**	**LIME**
Human accuracy	90.35%	83.03%	84.13%
% Neutral	8.48%	12.22%	19.22%

[2]All qualitative examples presented in our work are randomly selected from the outputs of the model initialization with the best faithfulness.

are judged as “neutral / can’t decide”, because the selected key words are neutral, or because positive and negative words are comparable in quantity. We exclude these neutral examples when computing the average accuracy for a participant, but also record the proportion of such examples as a proxy measure for purity. The final accuracy is averaged over multiple participants and reported in Table 3. It shows that our explanations are perceived to be more informative and contain fewer neural features, thus being more meaningful for human users to decide.

4.6 Multi-class Explanation

As discussed in Section 1, our method can explain multiple target classes given by the matrix W. Whereas L2X and VIBI do not have this luxury, LIME does this inefficiently as independent runs are required to interpret a particular class. Our explainer produces class-specific subsets of features with a single forward pass: given a learned W_x, simply select the column j ($W_x^{:,j}$) corresponding to the target class to be explained. Table 4 presents examples for explaining the two classes with the highest prediction probabilities. Qualitatively, the features selected by AIM are distinctive to each class while those obtained from LIME tend to be generic and overlap with one another.

Table 4: Comparison of 10 most relevant key words selected by AIM and LIME with respect to different target classes in AG News dataset.

Original document	*tacoma weathers ERP and CRM 'perfect storm' problems with a $50 million-plus rollout of SAP's ERP, CRM and other business apps in the city of tacoma, wash., have generated a storm of end-user complaints, bad press and a call for an independent audit of the situation.*	
Explainer	**Topic**	**Key words**
AIM	Sci/Tech	*SAP, plus, rollout, apps, ERP, ERP, user, CRM, CRM, problems*
AIM	Business	*$50, business, million, call, weathers, complaints, audit, independent, wash, ,*
LIME	Sci/Tech	*storm, CRM, problems, the, million, $50, independent, city, perfect, plus*
LIME	Business	*business, rollout, of, storm, tacoma, ERP, SAP, wash, press, the*
Original document	*wrapup 1-careless Chelsea and Arsenal let victories slip league leaders Chelsea allowed Bolton Wanderers to recover from two goals down to force a 2-2 draw at Stamford bridge in one of two major surprises in the premier league on Saturday.*	
Explainer	**Topic**	**Key words**
AIM	Sports	*league, league, Arsenal, Chelsea, Chelsea, Bolton, draw, victories, 1, premier*
AIM	World	*leaders, bridge, allowed, premier, force, Saturday, wrapup, two, two, wanderers*
LIME	Sports	*league, 2, careless, Chelsea, at, Arsenal, a, Saturday, leaders, bridge*
LIME	World	*league, 2, careless, leaders, Chelsea, Arsenal, goals, bridge, in, let*

Table 5: Faithfulness of our variants under conventional inference. * Proposed method. Results in this column are those reported for **AIM**^{c} in Table 2.

Training	**Learnable Locality***	Heuristic Locality	Learnable Locality	Heuristic Locality
Inference	Blackbox-driven		Explainer-driven	
IMDB	$\mathbf{99.98\% \pm 0.02\%}$	$89.50\% \pm 1.71\%$	$87.42\% \pm 0.32\%$	$54.84\% \pm 0.59\%$
HateXplain	$\mathbf{96.06\% \pm 0.91\%}$	$94.26\% \pm 0.81\%$	$77.18\% \pm 0.80\%$	$73.44\% \pm 2.64\%$
AG News	$\mathbf{97.54\% \pm 0.19\%}$	$94.58\% \pm 0.64\%$	$91.44\% \pm 0.14\%$	$73.48\% \pm 0.31\%$

5 Ablation Study

Here we demonstrate that by learning local distributions, an explainer can yield better explanations than heuristically sampling local instances as done by LIME. To mimic LIME, we first exclude the selection network $\mathcal{S}$ and approximator $\mathcal{G}$. We draw non-zero elements of z uniformly at random, with the number of such elements is also uniformly sampled. The explainer $\mathcal{E}$ is trained globally on these samples, denoted as $z_{\text{unif}} \in \{0,1\}^d$; this is done by minimizing $\mathbb{E}_x\Big[\text{CE}\left(\bar{y}_m, \sigma(W_x^T z_{\text{unif}})\right)\Big]$ - a modified version of $\mathcal{L}_1$ in 1 where z_{unif} is used in replacement of π_x. We name this training technique as **Heuristic Locality**, in comparison with the proposed method **Learnable Locality**. We also experiment with the strategy of inferring features according to the explainer, instead of utilizing black-box as proposed. Specifically, we select the weight vector $W_x^{:,j}$ where $j = \text{argmax}_c\, \sigma(W_x^T \pi_x)$. We apply explainer-driven inference to both cases of Learnable and Heuristic Locality; for the latter we again substitute π_x with z_{unif}. Table 5 highlights the importance of learning local distributions in order to mitigate the risk of having ill-conditioned local instances, as the model becomes less faithful when trained on heuristic perturbations. As discussed in Section 4.4, the explainer-driven strategy subjects the model to inconsistencies between the explainer and the black-box, which causes a degradation in faithfulness. Particularly combining with heuristic locality leads to a huge drop in performance of more than 20%.

6 Conclusion

We developed AIM - a novel model interpretation framework that integrates local additivity with instance-wise feature selection. The approach focuses on learning attributions across the target output space, based on which to derive features maximally faithful to the black-box model being explained. We provide empirical evidence proving the quality of our explanations: compact yet comprehensive, distinctive to each decision class and useful to human users.

References

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015; pp 1721–1730.

Rich, M. L. Machine learning, automated suspicion algorithms, and the fourth amendment. *University of Pennsylvania Law Review* **2016**, 871–929.

Nguyen, V.; Le, T.; De Vel, O.; Montague, P.; Grundy, J.; Phung, D. Information-theoretic Source Code Vulnerability Highlighting. 2021 International Joint Conference on Neural Networks (IJCNN). 2021; pp 1–8.

Lipton, Z. C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.

Mothilal, R. K.; Sharma, A.; Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020; pp 607–617.

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

Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Workshop at International Conference on Learning Representations. 2014.

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

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* **2014**,

Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. International conference on machine learning. 2017; pp 3145–3153.

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.-R. How to explain individual classification decisions. *The Journal of Machine Learning Research* **2010**, *11*, 1803–1831.

Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.

Chen, J.; Song, L.; Wainwright, M.; Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. International Conference on Machine Learning. 2018; pp 883–892.

Bang, S.; Xie, P.; Lee, H.; Wu, W.; Xing, E. Explaining a black-box using deep variational information bottleneck approach. *arXiv preprint arXiv:1902.06918* **2019**,

Slack, D.; Hilgard, S.; Singh, S.; Lakkaraju, H. How much should i trust you? modeling uncertainty of black box explanations. *arXiv preprint arXiv:2008.05030* **2020**, 6–25.

Situ, X.; Zukerman, I.; Paris, C.; Maruf, S.; Haffari, G. Learning to explain: Generating stable explanations fast. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021; pp 5340–5355.

Carter, B.; Mueller, J.; Jain, S.; Gifford, D. What made you do this? Understanding black-box decisions with sufficient input subsets. Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. 2019; pp 567–576.

Schwab, P.; Karlen, W. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems* **2019**, *32*.

Mahajan, D.; Tan, C.; Sharma, A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277* **2019**,

Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* **2016**,

Maddison, C. J.; Mnih, A.; Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* **2016**,

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **1998**, *86*, 2278–2324.

Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. 2017.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; Potts, C. Learning word vectors for sentiment analysis. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011; pp 142–150.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; Mukherjee, A. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *35*, 14867–14875.

Gers, F. A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural computation* **2000**, *12*, 2451–2471.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems* **2015**, *28*.

LeCun, Y.; Bengio, Y., et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **1995**, *3361*, 1995.

Jacovi, A.; Goldberg, Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020; pp 4198–4205.

Miller, G. A. WordNet: a lexical database for English. *Communications of the ACM* **1995**, *38*, 39–41.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **2014**, *15*, 1929–1958.

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

Appendix

A Experimental statistics

This section details data splits and hyperparameters used in our experiments for 3 text classification (IMDB / HateXplain / AG News) and 2 image recognition tasks (MNIST / Fashion-MNIST). α and β are the balancing coefficients on the loss terms in the final training objective. They are tuned via grid search for every dataset and the best ones are reported. These settings are applied consistently on the baseline models as well.

Table 6: Dataset statistics and hyperparameters.

Dataset	**Train/Dev/Test**	**No. of features**	α	β
IMDB	25000/20000/5000	400	1.8	1e−3
HateXplain	15000/4119/1029	200	0.1	1e−3
AG News	120000/6080/1520	400	0.1	1e−4
MNIST	14000/4623/3147	16	0.5	1e−3
Fashion-MNIST	15000/3000/3000	16	0.5	1e−3

B Model Efficiency

In this section, we analyze the trade-off between faithfulness and CPU times of our adaptive inference strategy, compared with LIME (trained at $K = 10$). We run adaptive inference sequentially on each instance to make it comparable with LIME, although it can be implemented much faster in mini-batch mode. Results are averaged over 5 random model initializations.

As K_{max} is set at higher values, our model explainer AIM^{a} achieves higher faithfulness with CPU times remaining relatively stable. At $K_{max} = 1$, our faithfulness already reaches above 97%. As for LIME, the model processes training and inference together for each instance, which is itself timely inefficient. To gain higher faithfulness requires more sampling perturbations at the cost of exponential increases in processing times.

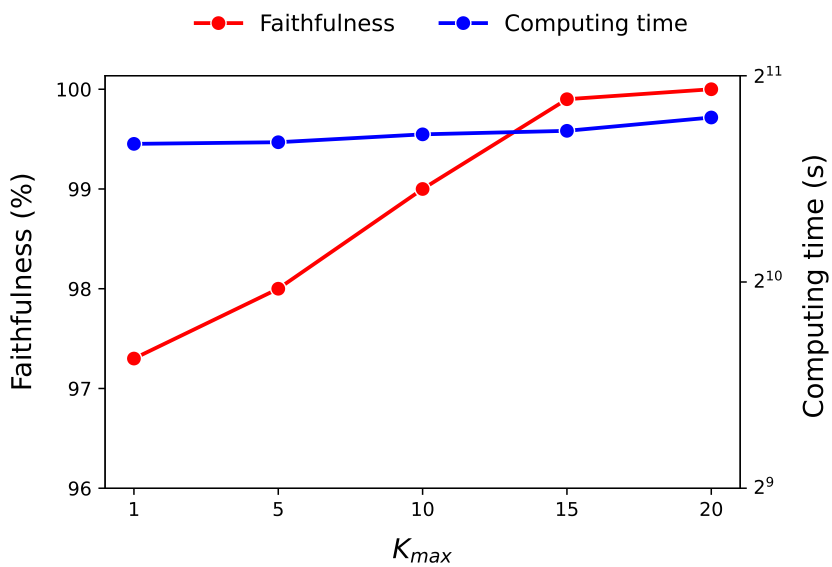

Figure 3: Faithfulness of AIM^a at various K_{max} vs. average time for adaptive inference on a single CPU and over 5000 test samples in IMDB dataset. Inference is done after training and sequentially for each instance.

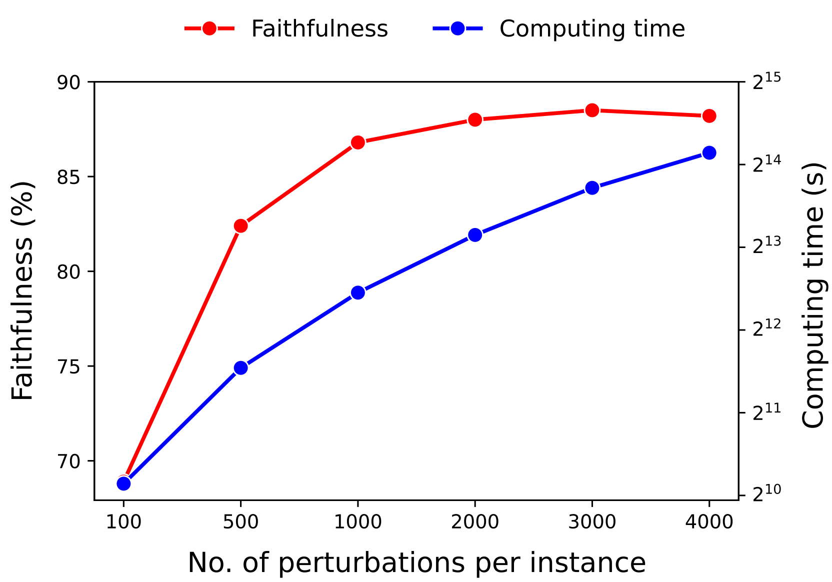

Figure 4: Faithfulness of LIME at $K = 10$ vs. number of random perturbations and average processing time on a single CPU and over 5000 test samples in IMDB dataset. LIME does training and inference together and sequentially for each instance.

C Training explainer $\mathcal{E}$ with Gumbel-Softmax relaxation

Here we analyze the performance of AIM when training the explainer on the relaxation of the Bernoulli variables z_x, instead of directly using the probabilities vector π_x. The first loss term $\mathcal{L}_1$ in Equation (1) is modified as follows

$$\mathcal{L}_1 = \mathbb{E}_x\Big[\mathrm{CE}\left(\bar{y}_m, \sigma(W_x^T \tilde{z}_x)\right)\Big]$$

Similar to $\mathcal{L}_2$, $\tilde{z}_x$ is obtained by applying Gumbel-Softmax trick on z_x: each element $\tilde{z}_x^i$ is sampled from the Concrete distribution as $\left[\tilde{z}_x^i, 1-\tilde{z}_x^i\right] \sim \mathrm{Concrete}(\pi_x^i, 1-\pi_x^i)$:

$$\tilde{z}_x^i = \frac{\exp\{\left(\log \pi_x^i + G_{i1}\right)/\tau\}}{\exp\{(\log(1-\pi_x^i) + G_{i0})/\tau\} + \exp\{(\log \pi_x^i + G_{i1})/\tau)\}}$$

with random noises G_{i0} and G_{i1} independently drawn from **Gumbel** distribution $G_t = -\log(-\log u_t),\ u_t \sim$ **Uniform**$(0,1)$. The other terms in the final training objective remains the same.

Table 7 reports the performance of our model using this method, compared with the originally proposed one. We find that during training, the sampling randomness requires longer training time for the model to converge and causes inconsistencies across datasets. This leads to higher variance and larger gap in performance between two inference techniques. Though the quality of the selected features is not severely impacted, faithfulness decreases up to nearly 10%. It signals that updating $\mathcal{E}$ with relaxed inputs sometimes makes it difficult to balance between faithfulness and compactness in the representations.

Table 7: Performance of AIM under adaptive (a) and conventional inference (c) when training the explainer on relaxed samples. Results are averaged over 5 random model initializations. * Proposed method. Results in these columns are taken from the main paper for comparison.

Explainer	**AIM**^{a}	**AIM**^{c}	**AIM**^{a}	**AIM**^{c}
Method	Without relaxation* (input π_x)		With relaxation (input $\tilde{z}_x$)	
		IMDB		
Mean K	1.15	10	2.06	10
Purity ↓	$3.23\% \pm 0.28\%$	$9.66\% \pm 0.72\%$	$\mathbf{1.28\% \pm 0.03\%}$	$\underline{4.12\% \pm 0.60\%}$
Brevity ↓	$\mathbf{1.03 \pm 0.00}$	$\underline{2.06 \pm 0.03}$	1.23 ± 0.10	2.59 ± 0.08
Faithfulness ↑	$\mathbf{99.98\% \pm 0.02\%}$	$\underline{99.98\% \pm 0.02\%}$	$97.06\% \pm 1.81\%$	$90.32\% \pm 5.72\%$
		HateXplain		
Mean K	1.85	10	2.71	10
Purity ↓	$\mathbf{6.76\% \pm 0.86\%}$	$25.27\% \pm 1.60\%$	$9.10\% \pm 1.73\%$	$\underline{22.07\% \pm 1.86\%}$
Brevity ↓	$\mathbf{1.34 \pm 0.05}$	4.16 ± 0.10	1.42 ± 0.03	$\underline{3.38 \pm 0.07}$
Faithfulness ↑	$\mathbf{98.86\% \pm 0.32\%}$	$\underline{96.06\% \pm 0.91\%}$	$94.40\% \pm 0.74\%$	$87.40\% \pm 1.86\%$
		AG News		
Mean K	1.56	10	2.07	10
Purity ↓	$\mathbf{0.64\% \pm 0.10\%}$	$\underline{4.47\% \pm 0.33\%}$	$0.91\% \pm 0.25\%$	$4.72\% \pm 0.35\%$
Brevity ↓	$\mathbf{1.14 \pm 0.01}$	$\underline{3.35 \pm 0.01}$	1.28 ± 0.14	3.37 ± 0.03
Faithfulness ↑	$\mathbf{99.48\% \pm 0.10\%}$	$\underline{97.54\% \pm 0.19\%}$	$97.52\% \pm 2.08\%$	$93.02\% \pm 4.81\%$

D Stability

One desirable property of a good model explainer is Stability (Situ et al., 2021) - that is the ability to produce the same explanations given similar examples. In the context of text explanations, the subsets of selected important words are expected to overlap in large quantity for two similar documents. We evaluate explanation stability through a simplified implementation of the measure *Intersection over Union* **(IoU)** originally proposed in Situ et al. (2021).

Given a specific instance x in the test set, we first search for the nearest neighbors $\mathcal{N}(x)$. The neighboring documents are defined to (1) have the same (black-box predicted) label and (2) either lexically or semantically similar. We adopt the ratio of overlapping tokens as a proxy metric for lexical similarity. Semantic similarity is measured via cosine similarity of their BERT representations, obtained by summing over the token representations of the last hidden state produced by a pre-trained BERT base (uncased), open sourced by Hugging Face (Wolf et al. 2019). We then select a set of 10 distinctive neighbors, consisting of top 5 semantically and top 5 lexically similar documents. Let v_x and $v_{x'}$ respectively denote the subsets of top K tokens selected for the instance x and its neighbor x', **IoU** is given as

$$\frac{1}{|\mathcal{N}(x)|} \sum_{x' \in \mathcal{N}(x)} \frac{|v_x \cap v_{x'}|}{|v_x \cup v_{x'}|}$$

To eliminate the effect of poor initialization, for each model explainer, we evaluate the model initialization with the highest faithfulness and compare the stability of top 10 explanations (results from conventional inference). Noticing that explainers sometimes favor a large number of stopwords, which may overestimate the measure, we exclude these tokens in the feature sets. Table 8 presents how each model performs on stability. By leveraging global information effectively, our model and L2X can pick up similar features better than LIME and VIBI. This again emphasizes LIME's limitation when only exploiting local information. Due to our black-box driven inference strategy, our explainer can achieve more stable explanations than L2X since we can avoid the inconsistent predictions between the explainer and the black-box.

Table 8: Stability of model explainers on top 10 explanations under conventional inference. Higher is better. Evaluation is done on the model initialization with the best faithfulness.

Explainer	**AIM**^{c}	**L2X**^{c}	**LIME**	**VIBI**^{c}
IMDB	**5.11%**	3.14%	2.28%	0.59%
HateXplain	**4.09%**	3.48%	3.57%	1.40%
AG News	**7.45%**	7.31%	5.04%	2.34%

E Qualitative Comparison

This section presents 12 additional qualitative examples to examine the quality of explanations of all model explainers. These examples are randomly selected from the outputs of the model initialization with the best faithfulness. Examples $9 - 12$ are particularly dedicated to illustrate multi-class explanations. Across all examples, we again demonstrate that our explanations are strongly consistent with black-box predictions, highly compact (by covering duplicates and synonyms) and distinctive to each decision class.

1. Original document: *this movie was a pleasant surprise for me. in all honesty, the previews looked horrible, up until the point where emma thompson and alan rickman appeared. so i rented it with reservation, but i thoroughly enjoyed this movie. it had great acting, a few good plot twists, and, of course, emma thompson and alan rickman. its definitely worth checking out.*

Truth: positive - **Model:** positive

Explainer	Key words
AIM	*great, enjoyed, definitely, worth, twists, surprise, thoroughly, me, good, reservation*
L2X	*enjoyed, pleasant, definitely, rented, horrible, surprise, worth, great, had, thoroughly*
LIME	*worth, great, enjoyed, and, it, checking, definitely, surprise, movie, thompson*
VIBI	*horrible, great, rickman, definitely, honesty, surprise, rented, reservation, the, acting*

2. Original document: *this may just be the worst movie ever produced. worst plot, worst acting, worst special effects...be prepared if you want to watch this. the only way to get enjoyment out of it is to light a match and burn the tape of it, knowing it will never fall into the hands of any sane person again.*

Truth: negative - **Model:** negative

Explainer	Key words
AIM	*worst, worst, worst, worst, plot, acting, any, just, hands, only*
L2X	*worst, worst, worst, worst, may, of, any, the, this, will*
LIME	*worst, special, of, acting, match, the, prepared, enjoyment, to, tapeX*
VIBI	*burn, worst, knowing, ,, ,, tape, the, the, special, ,*

3. Original document: *to me, "anatomie" is certainly one of the better movies i have seen. i dont think "anatomie" was primarily intended to be a horror movie but a movie questioning the ethics of science. if you watch it with that in mind, it turns into a really good film. the only annoying bit was the awful voice dubbing for the english version. how can you expect any non-german person to listen to these unbearable german accents for two hours? let native english speakers do the talking or use subtitles instead!!*

Truth: positive - **Model:** negative

Explainer	Key words
AIM	*awful, annoying, ?, unbearable, any, instead, "nt", talking, accents, or*
L2X	*awful, good, watch, instead, subtitles, unbearable, really, if, only, speakers*
LIME	*awful, can, non, seen, the, really, speakers, into, instead, annoying*
VIBI	*anatomie, questioning, anatomie, german, native, ethics, intended, instead, let, movie*

4. Original document: *i have seen this movie several times, it sure is one of the cheapest action flicks of the eighties. so, i think many viewers would definitely change the channel when they come across this one. but, if you are into great trash, "dragon hunt" is made for you. the main characters (the mcnamara twins) are sporting great moustaches and look so ridiculous in their camouflage dresses. one of the best scenes is when one of then gets shot in the leg and is still kicking his enemies into nirvana. this movie is really awful, but then again, it is a great party tape!*

Truth: negative - **Model:** positive

Explainer	Key words
AIM	*great, great, great, best, still, definitely, you, you, seen, action*
L2X	*best, ridiculous, great, still, cheapest, great, action, change, definitely, really*
LIME	*great, is, and, best, party, it, made, the, really, awful*
VIBI	*sporting, moustaches, mcnamara, nirvana, into, great, dragon, tape, great, party*

5. Original document: *this is not a competition against whites we are in this together if you want to criticise other whites this only makes it impossible to ever beat these spics and sand niggers from taking over just as the niggers have .*

Truth: hate/offensive speech - **Model:** hate/offensive speech

Explainer	Key words
AIM	*sand, niggers, niggers, spics, whites, whites, these, criticise, impossible, have*
L2X	*spics, niggers, beat, sand, these, niggers, ever, only, makes, and*
LIME	*niggers, this, whites, are, is, ever, have, makes, we, against*
VIBI	*criticise, this, ever, taking, just, these, from, want, we, a*

6. Original document: *all it does is boomerpost while pretending to be a slav and screeches anytime anyone says anything about jews like anyone other than a boomer would be retarded enough to believe slavs do not know their own 20 th century history at the hands of the chosenites .*

Truth: normal - **Model:** hate/offensive speech

Explainer	Key words
AIM	*slavs, boomer, jews, slav, screeches, anytime, retarded, than, their, the*
L2X	*retarded, boomer, would, anyone, anything, slav, history, says, enough, be*
LIME	*jews, believe, all, it, anytime, boomer, hands, boomerpost, retarded, 20*
VIBI	*anyone, anyone, history, century, than, while, believe, 20, be, would*

7. Original document: *peoplesoft chief threatens to sue over oracle statements peoplesofts chief executive accused oracle of spreading misleading information about his stock sales and threatened to sue for defamation .*

Truth: business - **Model:** business

Explainer	Key words
AIM	*stock, executive, sales, chief, chief, peoplesoft, peoplesoft, oracle, oracle, misleading*
L2X	*"s", chief, executive, peoplesoft, peoplesoft, stock, sales, sue, statements, accused*
LIME	*his, sales, oracle, chief, executive, about, accused, threatens, misleading, and*
VIBI	*defamation, peoplesoft, stock, oracle, executive, spreading, oracle, sue, misleading, threatens*

8. Original document: *blunkett gets tougher on drugs new police powers to prosecute offenders for possession if they test positive for drugs when they are arrested, even if the only drugs they have are in their bloodstream, are to be announced this week.*

Truth: world - **Model:** sports

Explainer	Key words
AIM	*test, positive, offenders, drugs, drugs, drugs, tougher, when, they, they*
L2X	*blunkett, police, positive, offenders, test, possession, week, they, only, [PAD]*
LIME	*test, offenders, they, tougher, drugs, the, for, gets, positive, are*
VIBI	*offenders, gets, to, positive, drugs, announced, are, are, „ ,*

9. Original document: *fda oks scientist publishing vioxx data (ap) ap - the food and drug administration has given a whistle-blower scientist permission to publish data indicating that as many as 139,000 people had heart attacks that may be linked to vioxx, the scientists lawyer said monday.*

Explainer	Topic	Key words
AIM	World	*attacks, people, lawyer, permission, linked, heart, ap, ap, whistle, indicating*
AIM	Sci/Tech	*scientist, scientist, scientist, data, data, may, many, be, ap, ap*
LIME	World	*ap, scientist, data, drug, people, monday, vioxx, attacks, s, may*
LIME	Sci/Tech	*monday, vioxx, food, drug, data, said, scientist, people, lawyer, had*

10. Original document: *tivo net loss widens; subscribers grow tivo inc.(tivo.o: quote, profile, research) , maker of digital television recorders, on monday said its quarterly net loss widened as it boosted spending to acquire customers, but subscribers to its fee-based tv service rose.*

Explainer	Topic	Key words
AIM	Business	*profile, quote, rose, quarterly, maker, widened, based, boosted, grow, fee*
AIM	Sports	*loss, loss, recorders, widened, television, monday, ;, subscribers, subscribers, but*
LIME	Business	*its, quote, digital, profile, widened, loss, said, monday, spending, customers*
LIME	Sports	*service, its, research, maker, profile, but, television, rose, monday, boosted*

11. Original document: *i could not believe the original rating i found when i looked up this film, 9.5? unfortunately it looks like i am not alone. the film, is slow and boring really, one of the sad things is that if the film had been given a realistic rating of around 5 or 6 then the expectation would not have been so high. unfortunately, this was not the case, so when watching the film, and seeing the poor story and acting, i am left giving it a 3/10 score. vinnie jones is superb in lock stock, and also snatch, and he plays a great hard man, however, he should stick to this role. its a bit like when stallone and schwarzenegger have done comedy films, they just don't work. neither can he play lead actor, he plays better as supporting or otherwise. when he plays lead, his acting talents are too 'in view' and shown up as not really very good. mean machine is another good example of this.*

Explainer	Sentiment	Key words
AIM	Negative	*poor, unfortunately, boring, 3/10, ?, otherwise, looks, acting, acting, looked*
AIM	Positive	*superb, great, also, very, bit, realistic, plays, plays, plays, supporting*
LIME	Negative	*poor, work, acting, given, lead, sad, one, rating, comedy, too*
LIME	Positive	*the, good, plays, when, not, unfortunately, acting, boring, then, case*

12. Original document: *the finest short i've ever seen. some commentators suggest it might have been lengthened, due to the density of insight it offers. there's irony in that comment and little merit. the acting is all up to noonan and he carries his thankless character perfectly. i might have preferred that the narrator be less "recognizable", but the gravitas lent is pitch perfect. this is a short for people who read, for those whose "bar" is set high and for those who recognize that living in a culture that celebrates stupidity and banality can forge contrary and bitter defenders of beauty. a beautiful short film. fwiw: i was pleased at the picasso reference, since i once believed that picasso was just another art whore with little talent; like, i assume, most people - until the day i saw some drawings he made when he was 12. picasso was a finer draftsman and a brilliant artist at that age than many artists will ever become in a lifetime. i understood immediately why he had to make the art he became known for.*

Explainer	Sentiment	Key words
AIM	Negative	*stupidity, talent, acting, suggest, just, been, banality, make, might, might*
AIM	Positive	*perfect, brilliant, perfectly, artists, beautiful, finest, pleased, celebrates, beauty, reference*
LIME	Negative	*and, he, ever, offers, a, it, short, defenders, in, saw*
LIME	Positive	*a, and, who, became, beautiful, he, is, short, the, i*

F Explaining image recognition models

This section describes the experiments for interpreting image recognition machine learning systems. The MNIST and Fashion-MNIST dataset respectively consist of 28×28 gray-scale images of handwritten digits and article clothing images. We train two simple neural networks on a subset of MNIST digits $0, 1, 2$ and Fashion-MNIST images of T-shirt/Trouser/Pullover. Both networks have the same architecture: 2 convolution layers kernel size 5 followed by 2 dense layers with output Softmax activation. The model on MNIST achieves 97.5% test accuracy while that on Fashion-MNIST gains 95.9%. We split each image into 4×4 patches size 7×7, resulting in a total of 16 features. Each input patch is first encoded by 2 convolution layers with kernel size 3, each followed by a ReLU activation and max pooling layer. The output module of the explainer $\mathcal{E}$ further consists of 2 top dense layers with ReLU activation and another dense layer at last. The selector $\mathcal{S}$ processes the encoded inputs through 2 dense layers, combined with Sigmoid activation layer to output probabilities. The approximator $\mathcal{G}$ simply flattens the inputs and feeds it to a dense layer on top of a classic Softmax final layer.

Table 9 reports faithfulness of our explanations on MNIST and Fashion-MNIST test sets under conventional inference, averaged over 5 random model initializations. Random selected examples of various scenarios are additionally presented for qualitative investigation. We find that the selected features are particularly useful to explain wrong decisions in terms of what spurious signals the black-box model relies on to make predictions e.g., the round shape to predict digit 0, or rectangular pattern at the bottom to predict a Trouser instead of a T-shirt.

Table 9: Faithfulness of AIM^c on MNIST and Fashion-MNST test set

K	5	8	10
MNIST	$92.67\% \pm 0.81\%$	$94.60\% \pm 0.52\%$	$95.08\% \pm 0.46\%$
Fashion-MNIST	$88.46\% \pm 0.33\%$	$92.23\% \pm 0.12\%$	$92.42\% \pm 0.14\%$

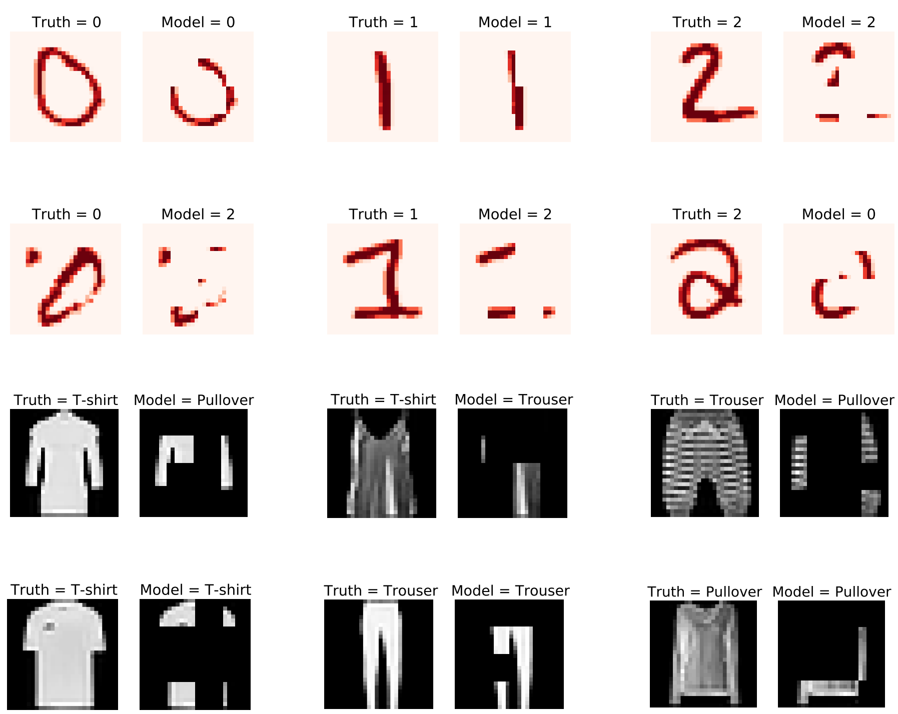

Figure 5: Explanations of the black-box model's predictions based on top 5 most relevant features.

G Human Evaluation

We ask 30 university students to infer the sentiments of 50 IMDB movie reviews, each of which is given only 10 key words obtained from each explainer under conventional inference. Each participant is presented with 3 sections, containing output examples from AIM, L2X and LIME respectively. Each section displays 50 sets of 10 key words corresponding to 50 different movie reviews. The information on which section belongs to which method is hidden and the ordering of examples within a section is randomized.

Movie Reviews Sentiment Evaluation

The survey has 3 sections. In each section, you will be given 50 lines of texts. Each line contains a few words extracted from a movie review. Given only such words, please infer the original sentiment of the movie review (Positive, Negative, Neutral / Can't decide).

Switch account

* Required

Based on the following words only, guess the original sentiment of the movie review *

	Positive	Negative	Neutral / Can't decide
1. [bad, poorly, total, saving, needless, whatever, far, not, not, only]	◯	◯	◯
2. [ridiculous, supposed, plot, plot, no, positive, anything, could, any, any]	◯	◯	◯
3. [awful, bad, bad, bad, horrible, grade, turkey, instead, reason, dialog]	◯	◯	◯

Figure 6: Human Evaluation Interface

H Ablation Study on L2X

H.1 L2X under adaptive inference

This section investigates the adaptive versions of L2X: we directly obtain the output weight vector as it is (since it is is single-class) and apply adaptive inference accordingly. Similar to our model, adaptive strategy impressively enhances the performance of L2X, reported in Table 10. No more than 3 features are required to approximate the black-box's decision for an instance on average. This number is far below the pre-chosen value K to train the explainer, which guarantees adequate computational efficiency and again supports our argument that fixating the explainer on a certain top K is sub-optimal and inflexible. This substantiates that optimizing K for each instance is essential since the performance deteriorates as we switch towards the conventional method and select more features than needed. In this regard, our proposed adaptive strategy is expected to work optimally on any instance-wise frameworks that produce a distribution of feature importance.

Table 10: Performance of L2X (trained at $K = 10$) under adaptive and conventional inference. $K_{max} = 10$ is chosen accordingly. Performance of AIM under adaptive inference is included in the last column for comparison. Results are averaged over 5 random model initializations.

Explainer	**L2X**^{c}	**L2X**^{a}	**AIM**^{a}
		IMDB	
Mean K	10	2.34	1.15
Purity $\downarrow$	$15.24\% \pm 1.04\%$	$6.45\% \pm 0.68\%$	$3.23\% \pm 0.28\%$
Brevity $\downarrow$	2.84 ± 0.04	1.31 ± 0.01	1.03 ± 0.00
Faithfulness $\uparrow$	$85.97\% \pm 0.25\%$	$94.92\% \pm 0.18\%$	$99.98\% \pm 0.02\%$
		HateXplain	
Mean K	10	2.81	1.85
Purity $\downarrow$	$22.89\% \pm 0.65\%$	$10.54\% \pm 1.17\%$	$6.76\% \pm 0.86\%$
Brevity $\downarrow$	4.41 ± 0.06	1.73 ± 0.03	1.34 ± 0.05
Faithfulness $\uparrow$	$74.05\% \pm 0.40\%$	$92.94\% \pm 0.49\%$	$98.86\% \pm 0.32\%$
		AG News	
Mean K	10	2.42	1.56
Purity $\downarrow$	$5.87\% \pm 0.45\%$	$2.33\% \pm 0.15\%$	$0.64\% \pm 0.10\%$
Brevity $\downarrow$	4.03 ± 0.05	1.51 ± 0.03	1.14 ± 0.01
Faithfulness $\uparrow$	$88.75\% \pm 0.24\%$	$96.30\% \pm 0.13\%$	$99.48\% \pm 0.10\%$

H.2 Inconsistency across K

The following table reports the performance of L2X trained at different values of K. It highlights the fact that a careful choice of K as a hyperparameter is crucial, and larger K does not necessarily yield better results. This is undesirable since in fact, larger K increases the chance of selecting meaningless features (lower purity) while does not guarantee faithfulness will go up accordingly. We also provide qualitative examples showing inconsistencies for selecting top K features i.e, the rankings of features vary across settings. For instance, though qualitatively considered an important feature, the word *amazing* in example 1 is selected in top 5 but does not appear in top 10 and and even ranks ninth in top 20. The same pattern is observed across examples.

Table 11: Performance of L2X when trained at 3 values of K for all datasets. Results are reported under conventional inference and averaged over 5 random model initializations.

L2X^c	$K = 5$	$K = 10$	$K = 20$
		IMDB	
Purity $\downarrow$	$10.48\% \pm 0.74\%$	$15.24\% \pm 1.04\%$	$18.33\% \pm 0.48\%$
Brevity $\downarrow$	1.70 ± 0.03	2.84 ± 0.04	5.38 ± 0.10
Faithfulness $\uparrow$	$82.30\% \pm 0.21\%$	$85.97\% \pm 0.25\%$	$84.28\% \pm 0.39\%$
		HateXplain	
Purity $\downarrow$	$19.40\% \pm 0.86\%$	$22.89\% \pm 0.65\%$	$33.24\% \pm 0.44\%$
Brevity $\downarrow$	2.81 ± 0.02	4.41 ± 0.06	7.83 ± 0.04
Faithfulness $\uparrow$	$68.48\% \pm 0.60\%$	$74.05\% \pm 0.40\%$	$77.08\% \pm 0.41\%$
		AG News	
Purity $\downarrow$	$1.87\% \pm 0.11\%$	$5.87\% \pm 0.45\%$	$18.55\% \pm 0.77\%$
Brevity $\downarrow$	2.47 ± 0.03	4.03 ± 0.05	7.38 ± 0.13
Faithfulness $\uparrow$	$84.02\% \pm 0.62\%$	$88.75\% \pm 0.24\%$	$82.20\% \pm 0.47\%$

1. Original document: *i saw Riverdance - the new show and loved it from the very first moment! it is an energetic tribute to Irish dance filled with brilliant dancing, music and choreography! the leads, Jean Butler and Colin Dunne had me captivated with their exquisite dancing! may they always keep shining and keep dancing. their on stage chemistry was amazing, and the unity between them on stage was obvious. they look like they were made to dance with each other! this show is my absolute favourite, and probably always will be. long live Riverdance!*

Truth: positive - **Model:** positive

K	**Key words**
5	*loved, favourite, energetic, brilliant, amazing*
10	*energetic, loved, favourite, tribute, brilliant, first, moment, dancing, butler, obvious*
20	*energetic, loved, butler, favourite, choreography, !, dancing, moment, amazing, other, dancing, captivated, brilliant, probably, Riverdance, is, ., tribute, music, !*

2. Original document: *how much longer will the west continue to put up with all of this shit from retarded third worlders and liberal cucks before we revert to a frontier mentality and just start the real violence .*

Truth: hate/offensive speech - **Model:** hate/offensive speech

K	**Key words**
5	*retarded, third, cucks, frontier, violence*
10	*third, retarded, cucks, worlders, frontier, mentality, violence, liberal, shit, and*
20	*worlders, third, retarded, and, frontier, mentality, to, from, will, longer, put, a, liberal, the, continue, shit, west, and, much, up*

3. Original document: *Yahoo and SBC extend partnership and plan new services Yahoo and SBC communications have agreed to collaborate to extend some of the online services and content they currently provide to PC users to mobile phones and home entertainment devices.*

Truth: sci/tech - **Model:** sci/tech

K	**Key words**
5	*Yahoo, Yahoo, users, phones, online*
10	*phones, online, Yahoo, users, Yahoo, devices, communications, mobile, collaborate, agreed*
20	*services, entertainment, collaborate, content, extend, and, and, PC, services, Yahoo, extend, some, have, partnership, provide, agreed, home, phones, the, and*

4. Original document: *riding high on the success of "rebel without a cause", came a tidal wave of teen movies. arguably this is one of the best. a very young Mcarthur excels here as the not really too troubled teen. the story concentrates more on perceptions of delinquency, than any traumatic occurrence. the supporting cast is memorable, Frankenheimer directs like an old pro. just a story of a young man that finds others take his actions much too seriously.*

Truth: positive - **Model:** positive

K	**Key words**
5	*best, memorable, one, troubled, the*
10	*best, memorable, too, teen, more, ,, riding, high, seriously, young*
20	*memorable, best, [PAD], very, ., one, high, seriously, troubled, any, perceptions, story, teen, is, young, ., directs, on, traumatic, success*

5. Original document: *i thought i was going to watch another friday the 13th or a halloween rip off, but i was surprised, its about 3 psycho kids who kill, theres not too many movies like that, i can think of mikey, children of the corn and a few others, its not the greatest horror movie but its a least worth a rent.*

Truth: negative - **Model:** negative

K	**Key words**
5	*surprised, least, 3, greatest, about*
10	*not, least, surprised, rent, ., i, was, [PAD], thought, worth*
20	*surprised, least, [PAD], think, halloween, worth, going, ., it, rent, was, mikey, children, a, about, ,, but, "s", [PAD], "s"*