TRANSFERLAB

appliedAI Seminar XAI

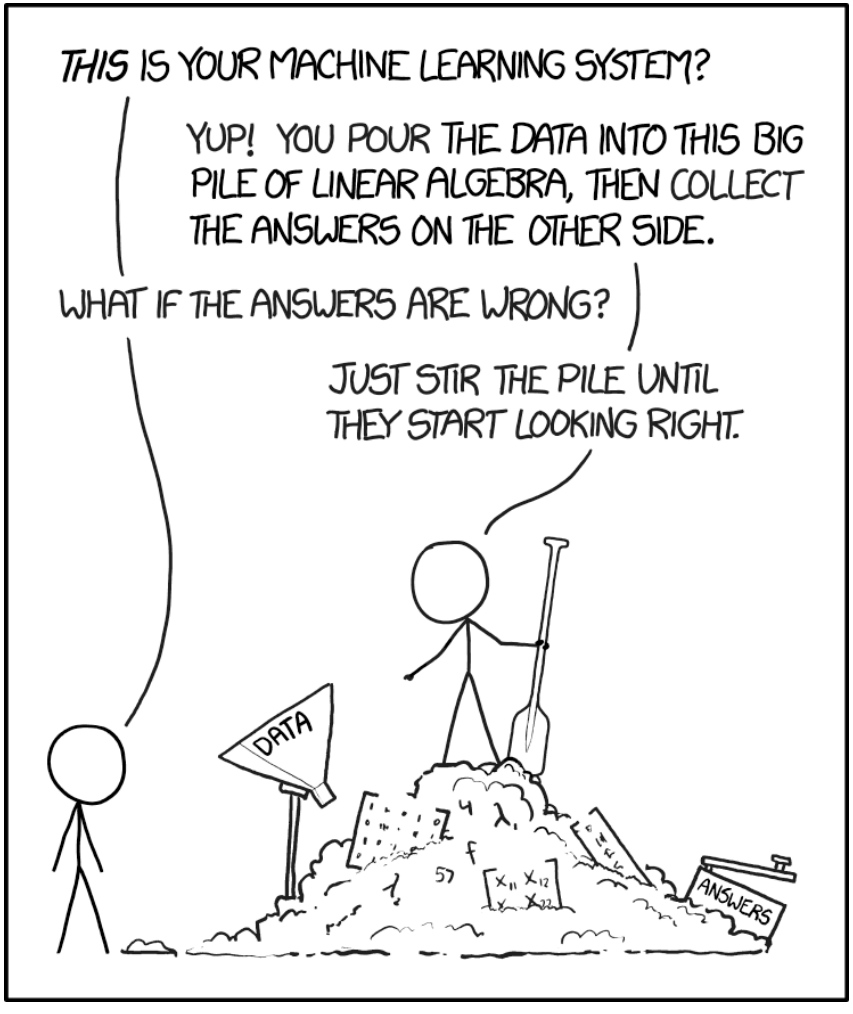
Learning to Explain: An Information-Theoretic Perspective on Model Interpretation (Chen et. al, 2018)

Kristof Schröder

12.10.2023

What are Black-Box Classifiers?

- Complex models whose inner workings are not directly interpretable.
- Often associated with deep learning, ensemble methods.
- Powerful in performance, but challenging in transparency.



https://xkcd.com/1838/

Why Explain Black-Box Classifiers?

- Trustworthiness: Can we trust what we don't fully understand? Can we understand, which features are relevant for an individual instance?
- Debugging: Identifying and rectifying model mistakes.
- Legal & Ethical: Meeting regulations and ethical standards.
- Stakeholder Communication: Explaining decisions to non-experts.

Figure 1: Cat or dog?[a]

[a] https://thedatafrog.com/en/articles/dogs-vs-cats/

What's out there in the wild?

Two popular examples for explainability methods:

- **LIME** (Local Interpretable Model-agnostic Explanations)[1]: LIME explains individual predictions by approximating black-box models locally with simpler interpretable models.
- **Kernel SHAP** (SHapley Additive exPlanations)[2]: Kernel SHAP assigns each feature an importance value based on average contributions across all possible feature combinations, grounded in game theory.

[1] Ribeiro et. al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016 [5]

[2] Lundberg, Scott and Lee, Su-In, A Unified Approach to Interpreting Model Predictions, 2017 [4]

What about Information Theory?

We look at the work of Chen et. al.[3], which tackles the problem from an information-theoretic point of view:

- Instance-wise feature selection for model interpretation, where each example's most informative features are identified.
- The feature selector is trained to boost the mutual information between chosen features and the outcome of the model in question.
- Variational formulation to allow for an efficient computation.

[3] Chen et. al, Learning to Explain: An Information-Theoretic Perspective on Model Interpretation, 2018[1]

Overview

A Primer on Information Theory

Information Theory

We start with revising some information-theoretic concepts. In the following, let X, Y be discrete random variables, with joint mass function $p_{(X,Y)}$ and marginal mass functions p_X, p_Y:

Information content / Surprisal:

- $I(y) = -\log P(Y = y)$
- an event with low probability has a high surprise
- If it rarely rains in the desert, the information content of *It will rain tomorrow* is very high because it's so unexpected.

Entropy:

- $H(Y) = -\sum_y P(Y = y) \log P(Y = y) = \mathbb{E}[I(Y)]$
- Measures the average surprise or uncertainty when observing random outcomes from Y.
- If a city's weather is truly unpredictable, with equal chances of *rain* or *no rain* on any given day, then the entropy of the weather forecast is high.

Information Theory

Conditional Entropy for specific realization:

- $H(Y|X=x) = -\sum_y P(Y=y|X=x) \log P(Y=y|X=x)$
- Calculates the uncertainty in one variable, given a specific known outcome of another variable.
- When you know it is raining today, how uncertain are you about the weather tomorrow?

Conditional Entropy:

- $H(Y|X) = -\sum_x p_X(x) H(Y|X=x) = \mathbb{E}[-\log P(Y|X)]$
- Average uncertainty of one variable when we have information about another variable.
- The conditional entropy gives the average uncertainty about tomorrow's rain, given different atmospheric pressures of today.

Information Theory

Mutual Information:

- $I(X;Y) = \sum_x \sum_y p_{(X,Y)}(x,y) \log \frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} = H(Y) - H(Y|X)$
- Reduction in uncertainty about Y after observing X
- If knowing today's atmospheric pressure greatly reduces your uncertainty about rain tomorrow, then the mutual information between pressure and rain is high.

Information Theory

To wrap up:

- Information content / Surprisal: $I(y) = -\log P(Y = y)$ (an event with low probability has a high surprise)
- Entropy: $H(Y) = \mathbb{E}[I(Y)]$ (Uncertainty about Y)
- Conditional entropy: $H(Y|X) = \mathbb{E}[-\log P(Y|X)]$ (Average uncertainty of Y when we have information about X)
- Mutual Information: $I(X; Y) = H(Y) - H(Y|X)$ (Reduction in uncertainty about Y after observing X)

Information-Theoretic Explanation

Setting

We consider classification problem with classes $[c] = \{1, \cdots c\}$, where the features are modeled by a random vector X (with realizations in $\mathbb{R}^d$) with marginal distribution

$$X \sim \mathbb{P}_X(\cdot),$$

and the predicted class Y by the classification model m is accessible via the family of conditional distributions:

$$(Y|x) \sim \mathbb{P}_m(\cdot|x), \quad x \in \mathbb{R}^d, \text{realization of } X$$

Caution: In this context, we are not discussing the population conditional class distributions. Our focus is solely on the conditional class distributions induced via the classification model.

Feature Importance

Taking a global perspective, one could ask which subset of the features is most relevant with respect to the mutual information to the target variable.[4] More concrete, consider the question of top-k important features. For the index set $[d] = \{1, \dots d\}$ define the admissible set

$$\mathcal{P}_k := \{S \subset 2^{[d]} || S| = k\},$$

i.e. all subsets of size k of the power set of the index set. Search for the optimal subset S^*

$$S^* := \arg\max_{S \in \mathcal{P}_k} I(X_S, Y),$$

where X_S is the restriction of X to a fix subset S.

[4] Gao et. al, Variational Information Maximization for Feature Selection, 2016[3]

Feature Importance

Solving the combinatorial problem

$$\arg\max_{S \in \mathcal{P}_k} I(X_S, Y)$$

is in general *NP*-hard. One possibility to tackle the problem, is to use greedy algorithms.[5] Besides the difficulties in the computation, this importance score is a global measure, i.e. it gives the most important features on average.

For complex models, we are more interested in a local importance, i.e. the most relevant features might vary, depending on the specific realization x.

[5] Das, Abhimanyu and Kempe, David, Submodular Meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection, 2011[2]

Instancewise Feature Selection

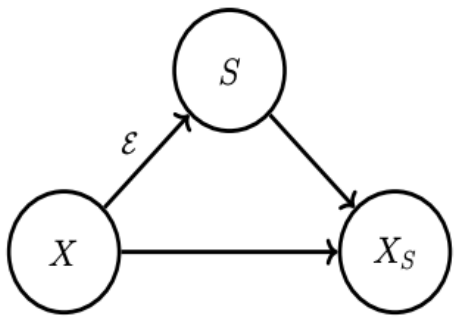


Figure 1: The graphical model of obtaining X_S from X.

An explainer $\mathcal{E}$ of size k is a mapping, which assigns a conditional distribution $\mathbb{P}(S|x)$ to every $x \in \mathbb{R}^d$. Given a subset $S = \mathcal{E}(x)$, we denote the sub-vector for the corresponding entries as x_S.

This defines a new random vector, which we again denote by $X_S \in \mathbb{R}^k$.

Instancewise Feature Selection

Using the definition of an explainer $\mathcal{E}$ and the constructed random vector X_S, we can formulate our objective for the instance-wise feature selection:

$$\max_{\mathcal{E}} I(X_S; Y) \quad \text{subject to} \quad S \sim \mathcal{E}(X)$$

So we want to maximize the mutual information between the response variable from the model and the selected features, as a function of the selection rule $\mathcal{E}$.

Variational Approximation

Mutual Information Objective

In order to derive a tractable formulation, let us inspect the objective:

$$\begin{aligned} I(X_S; Y) &= H(Y) - H(Y|X_S) \\ &= \mathbb{E}[\log \mathbb{P}_m(Y|X_S)] + H(Y) \end{aligned}$$

- $H(Y)$ is independent of S,
- maximizing the mutual information is equivalent to minimizing the conditional entropy,
- in order to get the conditional entropy, we would have to compute expectations under the conditional distribution $\mathbb{P}_m(\cdot|x_S)$, which is infeasible for general models.

Conditional Entropy

The (negative) conditional entropy can be expressed as:

$$-H(Y|X_S) = \mathbb{E}[\log \mathbb{P}_m(Y|X_S)] = \mathbb{E}_X \mathbb{E}_{S|X} \mathbb{E}_{Y|X_S}[\log \mathbb{P}_m(Y|X_S)]$$

In order to find a variational lower bound for the inner expression

$$\mathbb{E}_{Y|X_S}[\log \mathbb{P}_m(Y|X_S)],$$

we introduce (for the every fixed subset S) a conditional distribution $\mathbb{Q}_S(\cdot|x_S)$:

$$\mathbb{E}_{Y|X_S}[\log \mathbb{P}_m(Y|X_S)] = \underbrace{\mathbb{E}_{Y|X_S}\left[\log\left(\frac{\mathbb{P}_m(Y|X_S)}{\mathbb{Q}_S(Y|X_S)}\right)\right]}_{D_{KL}(\mathbb{P}_m(\cdot|x_S)||\mathbb{Q}_S(\cdot|x_S))\geq 0} + \mathbb{E}_{Y|X_S}[\log \mathbb{Q}_S(Y|X_S)]$$

$$\geq \mathbb{E}_{Y|X_S}[\log \mathbb{Q}_S(Y|X_S)]$$

Variational formulation

Let us define a collection of conditional distributions:

$$\mathbb{Q} = \{\mathbb{Q}_S(\cdot|x_s), S \in \mathcal{P}_k\}$$

with this, the maximization of the mutual information can be relaxed to:

$$\max_{\mathcal{E},\mathbb{Q}} \mathbb{E}[\log \mathbb{Q}_S(Y|X_S)] \quad \text{subject to} \quad S \sim \mathcal{E}(X)$$

Still, for generic $\mathbb{Q}, \mathcal{E}$ this is not tractable, so we have to restrict to suitable families.

Spoiler

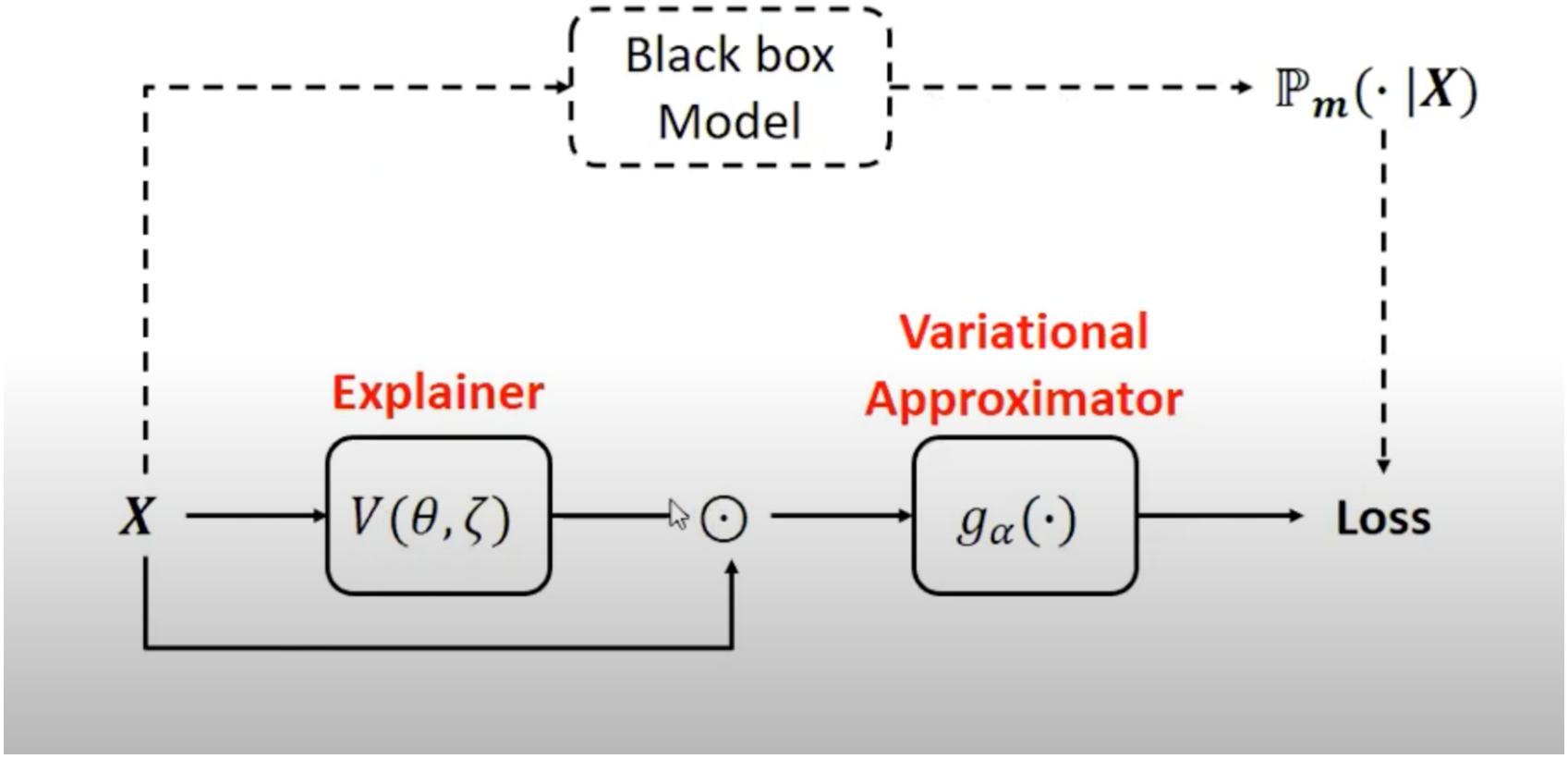


Figure 2: Diagram of training phase[6]

[6]Screenshot taken from `https://www.youtube.com/watch?v=id_CmUaTWpg`

Parametrizing the Variational Distributions

Criteria for $\mathbb{Q}$

- there should be suitable parametrization, which is accessible for optimization,
- the conditional distributions $\mathbb{Q}_S(\cdot|x_S)$ should be *close* to $\mathbb{P}_m(\cdot|x_S)$ (*small* KL-divergence)

The idea is to use a single neural network together with a masking operation. Using

$$g_\alpha : \mathbb{R}^d \times [c] \to [0,1],$$

where $[c] = \{1, \ldots, c\}$ are the possible classes, we define:

$$\mathbb{Q}_S(Y|x_S) := g_\alpha(\tilde{x}_s, Y),$$

with

$$(\tilde{x}_S)_i = \begin{array}{ll} x_i & i \in S \\ 0 & i \notin S \end{array}$$

Parametrizing the Variational Distributions

Using the neural network g_α, our maximization problem now looks like

$$\max_{\mathcal{E},\alpha} \mathbb{E}[\log g_\alpha(\tilde{X}_S, Y)] \quad \text{subject to} \quad S \sim \mathcal{E}(X)$$

where

$$\tilde{X}_S = Z_S \odot X \in \mathbb{R}^d,$$

and Z_S is the k-hot random vector (i.e. the mask), encoding the subset S.

- objective is differentiable in α (the parameters of the variational neural network),
- we still miss a smooth parametrization for $\mathcal{E}$

Smooth Relaxation of Masking

Given a categorical distribution represented by a one-hot vector, with category probabilities $p_1, \ldots, p_d$.

Gumbel-softmax re-parametrization trick (Concrete relaxation):

$$C_i = \frac{\exp\{(\log p_i + G_i)/\tau\}}{\sum_{j=1}^{d} \exp\{(\log p_j + G_j)/\tau\}},$$

with

$$G_i = -\log(-\log u_i), u_i \sim \text{Uniform}(0, 1)$$

We denote:

$$C \sim \text{Concrete}(\log p_1, \ldots, \log p_d)$$

The parameter τ is called temperature: the smaller the temperature, the closer the realizations of C resemble a one-hot vector.

Smooth Relaxation of Masking

To employ the Gumbel-softmax trick, we make the log probabilities learnable. We introduce a feature importance function

$$\omega_\theta : \mathbb{R}^d \to \mathbb{R}^d,$$

which should depend smoothly on θ and define k random vectors

$$C^j \sim \text{Concrete}(\omega_\theta(X)) \quad \text{i.i.d.} \qquad \text{for } j = 1, \ldots, k$$

and

$$V = (V_1, \ldots, V_d), \quad V_i = \max_j C_i^j$$

The random vector $V = V(\theta, \zeta)$ is a function of the parameters θ and auxiliary random variables

$$\zeta_i \sim \text{Gumbel}(0, 1) \quad \text{i.i.d.} \qquad \text{for } i = 1, \ldots, d$$

and we use this to smoothly approximate the k-hot vector Z_S.

Smooth Objective

To sum up, we replace the original objective

$$\mathbb{E}[\log \mathbb{P}_m(Y|X_S)]$$

with the parametrized relaxation

$$\begin{aligned}\textbf{Loss}(\theta, \alpha) &= \mathbb{E}_{X,Y,\zeta}[\log g_\alpha(V(\theta,\zeta) \odot X, Y)] \\ &= \mathbb{E}_{X,\zeta}\left[\sum_{y=1}^{c} \mathbb{P}_m(y|X) \log g_\alpha(V(\theta,\zeta) \odot X, y)\right]\end{aligned}$$

and ask to solve

$$\max_{\theta,\alpha} \textbf{Loss}(\theta, \alpha)$$

Training Phase

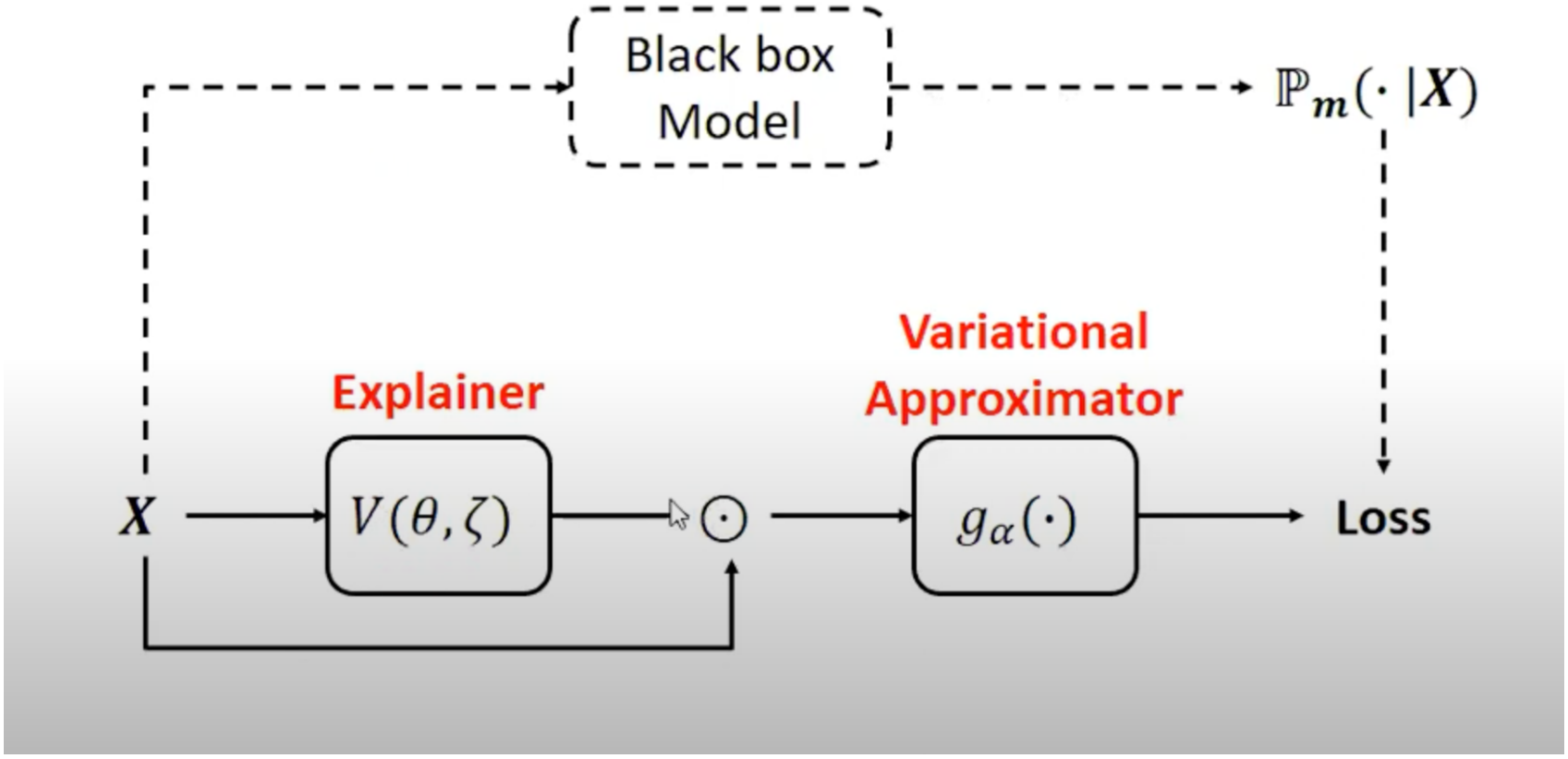


Figure 3: Diagram of training phase[7]

[7] Screenshot taken from `https://www.youtube.com/watch?v=id_CmUaTWpg`

Explanation Phase

Using the learned feature importance function ω_θ and a sample x:

- Compute the weights $\omega_\theta(x)$,
- select k features based on the top-k weights

The explanation phase requires one computation of ω_θ per sample.

Experiments

Sentiment Classification

The authors provide a binary sentiment classification problem for the Large Movie Review Dataset (IMDB). It consists of $50,000$ movie reviews, labeled either as positive or negative.

- IMDB-Word: explain a CNN with keywords
- IMDB-Sentence: explain a hierarchical LSTM with the most important sentence

Both models achieve an accuracy of around 90% on the test data. The feature importance networks (ω_θ) and variational approximators (g_α) are constructed from pre-trained word embeddings followed by convolutional, dense and pooling layers.[8]

[8] For conrete architectures, see Section 4.2 of [1]

IMDB-Word

Truth	Model	Key words
positive	positive	Ray Liotta and Tom Hulce shine in this sterling example of brotherly love and commitment. Hulce plays Dominick, (nicky) a mildly mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it, I promise, you'll be amazed at how such a wonderful film could go unnoticed.
negative	negative	Sorry to go against the flow but I thought this film was unrealistic, boring and way too long. I got tired of watching Gena Rowlands long arduous battle with herself and the crisis she was experiencing. Maybe the film has some cinematic value or represented an important step for the director but for pure entertainment value. I wish I would have skipped it.
negative	positive	This movie is chilling reminder of Bollywood being just a parasite of Hollywood. Bollywood also tends to feed on past blockbusters for furthering its industry. Vidhu Vinod Chopra made this movie with the reasoning that a cocktail mix of deewar and on the waterfront will bring home an oscar. It turned out to be rookie mistake. Even the idea of the title is inspired from the Elia Kazan classic. In the original, Brando is shown as raising doves as symbolism of peace. Bollywood must move out of Hollywoods shadow if it needs to be taken seriously.
positive	negative	When a small town is threatened by a child killer, a lady police officer goes after him by pretending to be his friend. As she becomes more and more emotionally involved with the murderer her psyche begins to take a beating causing her to lose focus on the job of catching the criminal. Not a film of high voltage excitement, but solid police work and a good depiction of the faulty mind of a psychotic loser.

Table 2. True labels and labels predicted by the model are in the first two columns. Key words picked by L2X are highlighted in yellow.

IMDB-Sentence

Truth	Predicted	Key sentence
positive	positive	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and excellent acting by all. Make some popcorn and have a great evening.
negative	negative	You get 5 writers together, have each write a different story with a different genre, and then you try to make one movie out of it. Its action, its adventure, its sci-fi, its western, its a mess. Sorry, but this movie absolutely stinks. 4.5 is giving it an awefully high rating. That said, its movies like this that make me think I could write movies, and I can barely write.
negative	positive	This movie is not the same as the 1954 version with Judy garland and James mason, and that is a shame because the 1954 version is, in my opinion, much better. I am not denying Barbra Streisand's talent at all. She is a good actress and brilliant singer. I am not acquainted with Kris Kristofferson's other work and therefore I can't pass judgment on it. However, this movie leaves much to be desired. It is paced slowly, it has gratuitous nudity and foul language, and can be very difficult to sit through. However, I am not a big fan of rock music, so its only natural that I would like the judy garland version better. See the 1976 film with Barbra and Kris, and judge for yourself.
positive	negative	The first time you see the second renaissance it may look boring. Look at it at least twice and definitely watch part 2. it will change your view of the matrix. Are the human people the ones who started the war? Is ai a bad thing?

Table 3. True labels and labels from the model are shown in the first two columns. Key sentences picked by L2X highlighted in yellow.

Evaluation

Post-hoc accuracy: On a test data set, run the explanation stage for every sample, mask the unselected features with zero padding and feed this into the original classification model. Compare this to the model output for full features.

Human accuracy: Provide humans with the feature subsets (i.e. top-10 Keywords or top-1 sentence), generated by the explainer, and ask them to give a prediction. Compare this prediction to the output of the classification model.[9]

	IMDB-Word	IMDB-Sent
Post-hoc accuracy	0.90.8	0.849
Human accuracy	0.844	0.774

[9] For a detailed description of the Amazon Mechanical Turk experiment, see [1, Section 4.2.1]

Discussion

Summary

- Instance-wise feature selection based on maximising mutual information.
- Tractable formulation using variational lower bound technique and the Gumbel-softmax trick.
- After initial training phase, explaining needs one forward pass per sample.
- Experiments to validate the explainer.

Questions

- Given a classifier model to explain, how to choose the feature importance function ω_θ and the variational approximator g_α?
- What are good strategies for choosing the temperature parameter τ in the Gumbel-softmax trick?
- What about considering other information-theoretic measures, e.g. Kullback-Leibler divergence[10]?
- Is post-hoc accuracy a good metric?

[10]Yoon, Jinsung and Jordon, James, INVASE: INSTANCE-WISE VARIABLE SELECTION USING NEURAL NETWORKS, 2019, [6]

Implementations

- Implementation of the authors using Tensorflow (research code): `https://github.com/Jianbo-Lab/L2X`
- Implementation in the OmniaXAI package using pytorch: `https://github.com/salesforce/OmniXAI`

Thank you!

References

J. Chen, L. Song, M. Wainwright, and M. Jordan.

Learning to Explain: An Information-Theoretic Perspective on Model Interpretation.

In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892. PMLR, July 2018.

A. Das and D. Kempe.

Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection.

In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1057–1064, Madison, WI, USA, 2011. Omnipress.

S. Gao, G. Ver Steeg, and A. Galstyan.

Variational Information Maximization for Feature Selection.

In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

S. M. Lundberg and S.-I. Lee.

A Unified Approach to Interpreting Model Predictions.

In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

M. Ribeiro, S. Singh, and C. Guestrin.

"Why Should I Trust You?": Explaining the Predictions of Any Classifier.

In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics.

J. Yoon, J. Jordon, and M. van der Schaar.

INVASE: Instance-wise variable selection using neural networks.

In *International Conference on Learning Representations*, 2019.