Policy-based Word Subset Selection for Explaining Black-Box Language Models

Minyoung Hwang*
Department of Artificial Intelligence
Korea University
Seoul, Korea
minyoung58@korea.ac.kr

Seokhyun Lee*
Department of Artificial Intelligence
Korea University
Seoul, Korea
nshuhsn@korea.ac.kr

Changhee Lee
Department of Artificial Intelligence
Korea University
Seoul, Korea
changheelee@korea.ac.kr

***Abstract*—The development of deep language models (DLMs) has made them more widely used. This has led to a growing need for tools that can help understand how these models work, especially when it comes to understanding the reasoning behind their outputs. This explainability is emerging as a key factor in building trust between users and technologies. Achieving meaningful interpretability remains a significant challenge, especially when DLMs are considered *black-box* systems, where internal details like parameters and gradients are inaccessible. Although many techniques have been proposed, most struggle to meet two critical goals simultaneously: (i) maintaining efficiency during inference, and (ii) remaining compatible with black-box models without causing out-of-distribution behaviors. To overcome these limitations, we introduce a method that explains model predictions by selecting a concise and informative subset of input words. Our approach involves training a lightweight selection network to identify a minimal yet informative subset of input tokens. Once trained, this network operates with high efficiency, directly identifying a salient subset of words as the explanation for new samples at inference time. For training, we leverage policy gradients for optimization, which critically allows our method to operate without requiring gradient information from the target DLM, thus making it inherently applicable to black-box systems by directly interacting with the target DLM solely through its input-output predictions without requiring any gradient information.**

***Index Terms*—explainable AI, black-box language model, word subset selection**

I. Introduction

Many of the most powerful Deep Language Models (DLMs) are now deployed as "black-box" services (e.g., APIs). While this paradigm accelerates adoption, it introduces a fundamental challenge: their internal decision-making processes are entirely opaque. When these models are applied in sensitive fields like medicine or law, this lack of transparency creates an urgent need for methods that can explain their predictive rationale without access to internal architecture, parameters, or gradients, which is essential for ensuring trustworthiness and accountability.

One of the main challenges of operationalising model explanations is the trade-off between performance and practicality. Instance-wise explanation methods [1, 2, 3, 4, 5, 6], while flexible, are often too computationally expensive for real-world use due to their need for repeated model access or per-instance optimization, precluding their use in low-latency environments. On the other hand, train-based approaches [7, 8, 9, 10], which train a separate explainer model for fast single-shot explanations, face a different challenge. In true black-box settings where gradients are inaccessible, they must often rely on training a surrogate predictor. However, accurately approximating a large-scale DLM with a surrogate is a prohibitively resource-intensive task, making this approach unscalable and impractical for many real-world applications.

Our method achieves both high efficiency and true black-box compatibility by framing explanation as a selection problem. We train a reusable selector network that operates in a single forward pass at test time, ensuring low-latency inference. This selector is optimized using policy gradients, which circumvents the need for internal model access. The selection for each word is modeled as a Bernoulli trial, and the policy is trained with rewards computed from the predictions of the black-box model on the selected subset of words alone.

**These authors contributed equally to this work.*

II. Related Works

A. Instance-wise Method

A significant body of work has focused on instance-wise methods, which generate an explanation for a single prediction at a time by analyzing the model's behavior concerning that specific input. These can be broadly categorized into perturbation-based and gradient-based approaches.

Perturbation-based methods locally approximate the model's decision boundary. LIME [6] trains a simple and interpretable model on perturbed samples in the vicinity of the instance being explained. SHAP [5], grounded in cooperative game theory, computes optimal feature importance values (Shapley values) and provides strong theoretical foundations. Although powerful, both LIME and SHAP typically require a large number of queries to the target model to generate a single explanation.

Gradient-based methods, such as Integrated Gradients (IG) [1] and DeepLIFT [3], calculate feature importance by propagating attribution scores from the output back to the input. These methods are often more efficient than perturbation-based approaches, but fundamentally require access to the model's gradients, making them inapplicable to the true black-box settings we address.

A common limitation across all these instance-wise methods is their computational cost at inference time. The need for multiple model queries or costly optimization for each new explanation makes them impractical for real-world scenarios that require low-latency or high-throughput analysis.

B. Learning-based Method

To overcome the efficiency bottleneck of per-instance methods, another line of research focuses on learning-based approaches. Pioneering works in this area, such as L2X (Learning to Explain) [9], introduced a paradigm where an explainer model and a local surrogate model (or predictor) are trained jointly. In this setup, the explainer learns to select a concise subset of features that are maximally informative for the surrogate, not the original target model. However, this approach raises significant questions about the faithfulness of the explanation. Since features are selected to explain the behavior of the surrogate, it is not guaranteed that the resulting explanation accurately reflects the true rationale of the original, more complex target model. Furthermore, this paradigm faces a significant scalability challenge. Training a surrogate model with sufficient capacity to accurately approximate a large-scale DLM requires substantial data and computational resources.

Other learning-based methods have different dependencies that limit their black-box applicability. For example, while the LTX [7] is presented as black-box compatible, its methodologies presuppose access to information beyond simple input-output pairs. Its initial step involves cloning the target model, which in itself requires full access to the model's architecture and weights, a condition unmet in true black-box scenarios. Furthermore, its optimization process depends on the gradient flow from this target model.

However, CXPlain [10] takes a different approach to avoid this dependency. It operates directly on the target model by training an explainer to predict pre-computed feature importance scores, calculated using a leave-one-out method. While this allows the explainer to learn without a surrogate or direct gradient access, it introduces other significant challenges, particularly for modern DLMs. Firstly, the complexity of the method increases with the number of features, making it difficult to scale effectively when each token in a long sequence is treated as an individual feature. Secondly, powerful DLMs are often highly robust to the removal of individual tokens, meaning the leave-one-out approach can result in uniform importance scores for all tokens. This makes it difficult to discern which features are genuinely critical, thereby limiting the utility of the resulting explanation. Therefore, despite their different strategies, existing learning-based approaches remain impractical for providing faithful and scalable explanations for large-scale black-box DLMs.

III. Problem Formulation

We consider the problem of explaining a pre-trained DLM, $g : X^T \to [0,1]^C$, which has been trained for a specific classification task. The model processes an input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T) \in X^T$, a sequence of T words from the input space X. A key consideration is that the sequence length, T, is variable and changes with each input. We treat g as a true black-box, assuming that while we can query the model to obtain its output $g(\mathbf{x})$, we have no access to its internal states, such as its parameters, gradients, or proprietary tokens (e.g., `[MASK]`, `[PAD]`).

The objective of our explanation is to identify the subset of words in $\mathbf{x}$ most responsible for its prediction. To this end, we formulate the problem as learning a binary selection mask $\mathbf{m} = (m_1, \ldots, m_T) \in \{0,1\}^T$, where $m_t = 1$ indicates that the t-th word is selected as part of the explanation.

Our formulation is predicated on the assumption that replacing non-essential words with a neutral placeholder does not significantly alter the model's decision-making process. We formalize this substitution as follows:

$$\tilde{\mathbf{x}} = \mathbf{m} \odot \mathbf{x} + (\mathbf{1} - \mathbf{m}) \cdot x_{\texttt{NULL}} \tag{1}$$

Here, the perturbed input $\tilde{\mathbf{x}}$ is constructed using an element-wise multiplication with the mask $\mathbf{m}$, while a placeholder word, $x_{\texttt{NULL}}$, fills the positions of the masked-out tokens. Therefore, our ultimate objective is to solve the following optimization problem:

$$\min_{\theta} \ \mathbb{E}_{\mathbf{x}\sim p_X} \mathbb{E}_{\mathbf{m}\sim \texttt{Bern}(\pi_\theta(\mathbf{x}))} \big[\ell\big(g(\mathbf{x}), g(\tilde{\mathbf{x}})\big) + \lambda ||\mathbf{m}||_0 \big]. \tag{2}$$

IV. Method

To address the challenges outlined above, we introduce a keyword-based explanation method, which we refer to as **PS2**, a novel framework for efficiently generating explanations for black-box DLMs. Our approach centers on training a reusable selector network. This network is optimized using a policy gradient strategy, which allows it to learn how to select informative subsets of words without requiring access to the target model's internal gradients. Crucially, this policy gradient approach utilizes hard, discrete sampling for word selection. This ensures the target model is always evaluated on realistic, in-distribution inputs (i.e., actual word subsets), avoiding the out-of-distribution issues common to methods that use continuous relaxations and thereby enhancing the faithfulness of the resulting explanation. The result is an explainer that is both fast at inference time and inherently compatible with black-box systems.

A. Selector Network Architecture

Our selector network, π_θ, is designed to be efficient, flexible, and context-aware. It consists of two main components: a fixed feature extractor and a trainable selection module.

First, to enable context-aware selections, we leverage a powerful, pre-trained DLM, $f : \mathcal{X}^T \to \mathbb{R}^{T\times d}$, as a feature extractor. For any given input $\mathbf{x}$ of length T, we compute contextualized embeddings for all tokens. Crucially, this feature extractor f is kept frozen during the training of our selector. This allows us to benefit from its rich linguistic knowledge without incurring high computational costs, ensuring our method has high training efficiency, as only the small selection module needs to be optimized.

Second, the selection module is implemented as a simple shared MLP. Specifically, we compute the selector output using the shared MLP architecture as follows:

$$\pi_\theta(\mathbf{x})_t = \sigma(\texttt{MLP}(f(\mathbf{x})_t)) \quad \text{for } t = 1, \ldots, T \tag{3}$$

where σ is a sigmoid function. This MLP processes the contextualized embedding of each word $f(\mathbf{x})_t$ independently to compute word-level selection probability $\pi_\theta(\mathbf{x})_t$. This resulting probability distribution, $\pi_\theta(\mathbf{x})$, is then used as the parameter for a Bernoulli distribution, from which a binary mask $\mathbf{m} \sim \texttt{Bern}(\pi_\theta(\mathbf{x}))$. Finally, this mask is then used, along with a neutral placeholder $\mathbf{x}_{\text{NULL}}$, to transform the original input $\mathbf{x}$ into the final sequence $\tilde{\mathbf{x}}$ according to Equation (1). A critical advantage of this design is its ability to naturally handle variable-length sequences, as applying the same MLP to each token allows our method to dynamically generate an explanation mask for any given input.

The neutral placeholder, $x_{\texttt{NULL}}$, is deliberately chosen to be a semantically neutral word (e.g., "the") that is unlikely to affect the model's original prediction. Performing substitution directly in the input text space, rather than manipulating special token IDs or embeddings, ensures our method is compatible with true black-box models where such internal interventions are forbidden. Furthermore, this approach guarantees that the model is always queried with natural in-distribution input, which helps to ensure a faithful explanation.

B. Optimizing the Selector via Policy Gradients

To ensure explanations are grounded in the actual task outcome, we formulate the loss term $\ell(g(\mathbf{x}), g(\tilde{\mathbf{x}}))$ from Equation (2) in a supervised manner. By incorporating the ground-truth label y, we guide the selector to identify subsets that are not just influential but also relevant to the correct classification.

The selector network is then trained using the REINFORCE algorithm [11], a classic policy gradient method. This approach is well suited for our task as it directly handles the discrete action space of word selection, thus circumventing the need for continuous approximations or relaxations [12, 13]. Adherence to discrete selections ensures that the target model is only ever evaluated on valid, in-distribution inputs. Furthermore, this gradient-free optimization allows the selector to learn using only the scalar reward signals obtained from querying the black-box model, requiring no access to the target model's internal gradients and making it perfectly suited for our target setting.

The policy gradient theorem provides an unbiased estimator for the gradient of our objective function $\mathcal{L}(\theta)$ from Equation (2) concerning the selector's parameters θ:

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}, y \sim p_{XY}} \Big[\mathbb{E}_{\mathbf{m} \sim \texttt{Bern}(\pi_\theta(\mathbf{x}))} \Big[\nabla_\theta \log p_\theta(\mathbf{m}) \cdot \big(\ell_y\big(g(\mathbf{x}), g(\tilde{\mathbf{x}})\big) + \lambda ||\mathbf{m}||_0 \big) \Big] \Big]. \tag{4}$$

The term $p_\theta(\mathbf{m}) = \prod_{t=1}^{T} (\pi_\theta(\mathbf{x}))_t^{m_t} (1 - (\pi_\theta(\mathbf{x}))_t)^{1-m_t}$ is the probability of sampling the gate vector $\mathbf{m}$. This corresponds to the probability mass function of a multivariate Bernoulli distribution, which is parameterized by the word-level selection probabilities output by our selector, $\pi_\theta(\mathbf{x})$.

V. Experiments

A. Experiment Setup

Datasets. We use the Movies dataset [14] for its binary sentiment labels and, more importantly, its human-annotated rationales. These rationales are provided as specific text spans within each review (a single review may contain multiple spans). The availability of these ground-truth annotations enables a precise, token-level assessment of explanation performance.

Performance Metrics. We evaluate the discriminative power of the selected word subsets using three standard metrics: classification accuracy (ACC), AUROC, and AUPRC. This evaluation is performed at multiple levels of sparsity, corresponding to gating rates of 5%, 10%, and 15%. The metrics are computed based on the black-box model's predictions when provided with only the selected words as input.

Benchmarks. We evaluate our method against two classes of representative baselines. The first class consists of per-instance methods that estimate importance without a separate training phase, including KernelSHAP [5], LIME [6], and Integrated Gradients (IG) [1]. The second class consists of learning-based methods, including L2X [9], LTX [7], and CXplain [10].

Among the learning-based methods, L2X can bypass the difficulty of training a high-capacity surrogate. To ensure a fair and rigorous comparison, our experimental setup grants it direct oracle access to the target DLM for training. Furthermore, to maintain a level playing field, all learning-based methods (L2X, LTX, and CXplain) were trained using the same word embeddings provided by $f(\mathbf{x})$. A key limitation, however, is that L2X, along with IG and LTX, relies on the target model's gradients. While this dependency places these methods outside our black-box definition, we include them to benchmark our method against such gradient-aware approaches.

Our experimental setup is as follows: We implement our selector network as a 2-layer MLP with ReLU activation functions. For the fixed feature extractor, we use a 12-layer transformer model and extract the contextualized embeddings from its 10th layer. We train the selector for 40 epochs using the Adam optimizer [15] with a mini-batch size of 32 for the Movies dataset. We perform a hyperparameter search for the learning rate over the set 0.0001, 0.00025, 0.0005, 0.001. For the policy gradient update, we estimate the gradient using 8 samples per input instance. The sparsity regularization coefficient λ in our objective function is set to 0.00625.

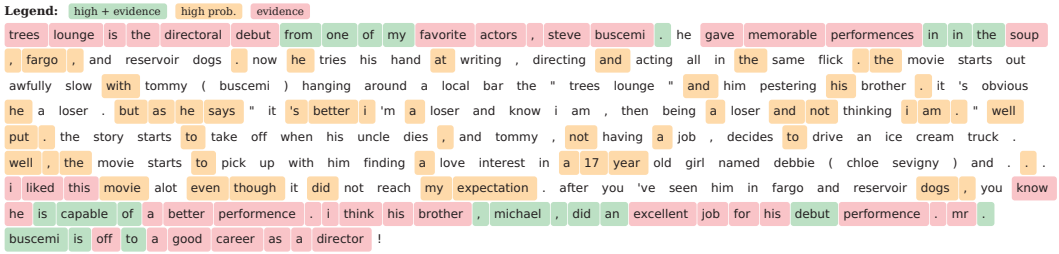
B. Experiment Result

We evaluate our proposed method both quantitatively and qualitatively against baseline approaches. The results demonstrate the superior performance of our method in identifying discriminative and faithful explanations.

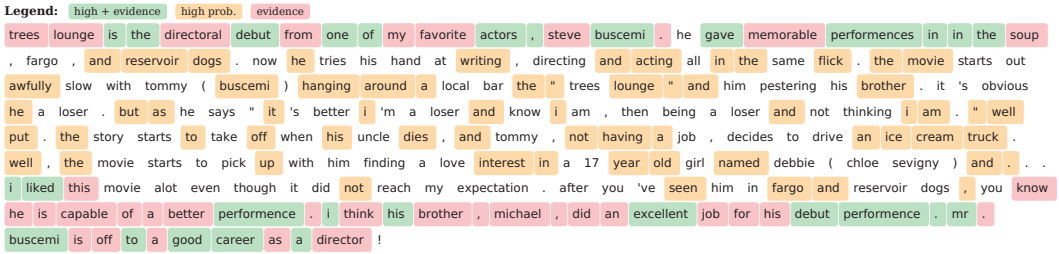
TABLE I: The Movies dataset: Performance comparison (ACC, AUROC, AUPRC) across selection rates (5%, 10%, 15%).

Method	5% ACC	5% AUROC	5% AUPRC	10% ACC	10% AUROC	10% AUPRC	15% ACC	15% AUROC	15% AUPRC
LIME	0.498 ± 0.043	0.501 ± 0.044	0.509 ± 0.029	0.496 ± 0.017	0.506 ± 0.020	0.528 ± 0.015	0.517 ± 0.019	0.518 ± 0.050	0.535 ± 0.052
SHAP	0.408 ± 0.070	0.416 ± 0.071	0.452 ± 0.055	0.474 ± 0.021	0.470 ± 0.025	0.497 ± 0.022	0.472 ± 0.032	0.471 ± 0.039	0.496 ± 0.022
IG	0.560 ± 0.067	0.416 ± 0.071	0.452 ± 0.055	0.475 ± 0.004	0.467 ± 0.015	0.488 ± 0.023	0.471 ± 0.016	0.447 ± 0.053	0.470 ± 0.057
CXPlain	0.522 ± 0.021	0.512 ± 0.025	0.526 ± 0.025	0.512 ± 0.004	0.571 ± 0.019	0.571 ± 0.010	0.524 ± 0.006	0.567 ± 0.014	0.582 ± 0.015
L2X*	0.518 ± 0.043	0.506 ± 0.092	0.541 ± 0.113	0.561 ± 0.054	0.596 ± 0.060	0.590 ± 0.080	0.583 ± 0.065	0.628 ± 0.071	0.622 ± 0.077
LTX*	0.560 ± 0.044	0.550 ± 0.074	0.560 ± 0.106	0.589 ± 0.066	0.605 ± 0.078	0.625 ± 0.104	0.599 ± 0.086	0.655 ± 0.088	0.664 ± 0.095
Ours	0.567 ± 0.029	0.761 ± 0.031	0.767 ± 0.030	0.733 ± 0.015	0.805 ± 0.016	0.797 ± 0.025	0.720 ± 0.027	0.816 ± 0.035	0.808 ± 0.038
Black Box (Full text)				**0.859**	**0.942**	**0.944**			

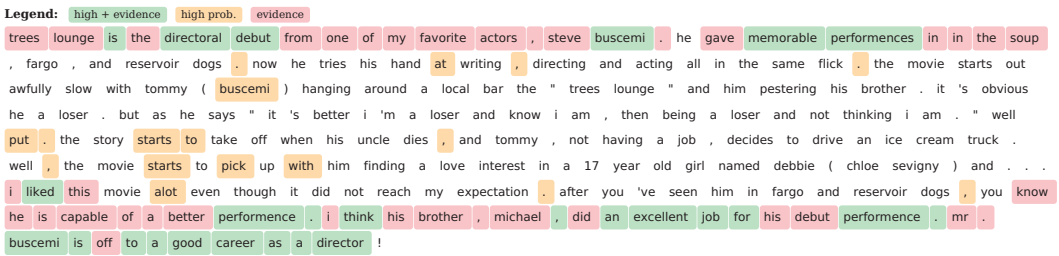
* Methods are adapted to our setting for fair comparison.



(a) LIME



(b) IG



(c) Our Method

Fig. 1: Qualitative examples of explanations generated by our proposed method on the Movies dataset, compared against ground-truth human rationales.

C. Quantitative Analysis

Table I presents the quantitative comparison of our method against the baselines across three selection rates (5%, 10%, and 15%). The results clearly show that our proposed method consistently and significantly outperforms all baseline explanation methods across all evaluation metrics (ACC, AUROC, and AUPRC) and at all sparsity levels.

For instance, at a 10% gating rate, our method achieves an AUROC of 0.805, a substantial improvement over the next best baseline, LTX of 0.605. This performance gap is consistent across all metrics. Notably, our method's performance gracefully improves as more features are selected, approaching the upper-bound performance of the full-text black-box model. In contrast, the performance of the baseline methods saturates more quickly or improves only marginally. This demonstrates the effectiveness of our policy-based selection strategy in

identifying highly discriminative word subsets that are faithful to the model's behavior.

D. Qualitative Analysis

In addition to the quantitative results, we provide a qualitative comparison in Fig.1 to visually assess the quality of the generated explanations. The figure displays a representative sample from the Movies dataset, showing the explanations generated by LIME, IG, and our method for the 15% selection rate setting. These are compared against the ground-truth human rationale to highlight the differences in alignment.

As the legend indicates, words highlighted in **green** represent correctly identified human rationales (true positives), words in **orange** are selected by a method but are not part of the ground-truth rationale (false positives), and words in **red** are human rationales missed by the method (false negatives).

It is visually evident that the explanations from LIME and IG are noisy, selecting several non-essential words (many orange highlights) while failing to capture the full extent of the human rationale (several red highlights). For example, It is visually evident that LIME's explanation is often incomplete, failing to capture some of the most critical parts of the human rationale. While IG identifies most of the important words within the human rationale, it suffers from low precision, highlighting numerous extraneous words that are not part of the ground truth. These false positives, such as the phrase "an ice cream truck," often lack a clear connection to the sentiment analysis task. In contrast, our method generates a more coherent and precise explanation that aligns much more closely with the human-annotated spans. It successfully identifies key phrases like "memorable performances", "an excellent job" and "a good career as a director" while selecting very few incorrect words. This visual evidence supports our quantitative findings, suggesting that our method learns to identify more faithful and human-interpretable rationales.

VI. Conclusion

We addressed the critical challenge of efficiently generating faithful explanations for true black-box Deep Language Models. We introduced PS2, a novel framework that trains a lightweight, reusable selector network to identify concise and informative word subsets as explanations. Our core contribution is the use of a policy gradient strategy, which optimizes the selector using only the target model's final output predictions. This approach circumvents the need for internal model access, such as gradients or architectural details.

Our extensive experiments demonstrated that our method significantly outperforms existing per-instance and learning-based baselines on both quantitative metrics and in qualitative alignment with human rationales. The results confirm that our method successfully achieves both high efficiency at inference time and true black-box compatibility without sacrificing the faithfulness of the explanations. By providing a practical and scalable solution, our work represents a meaningful step towards making the decisions of opaque language models more transparent and trustworthy.

There are several exciting directions in which future work could explore. One promising avenue is applying the proposed framework to other tasks, such as text generation or regression. Additionally, examining more sophisticated policy optimisation algorithms could enhance training stability and sample efficiency.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00358602) and by the Institute of Information and communications Technology Planning & Evaluation (IITP) funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University))

References

[1] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[2] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 2016, pp. 63–71.

[3] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMlR, 2017, pp. 3145–3153.

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] O. Barkan, Y. Asher, A. Eshel, Y. Elisha, and N. Koenigstein, "Learning to explain: A model-agnostic framework for explaining black box models," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 944–949.

[8] J. Yoon, J. Jordon, and M. van der Schaar, "INVASE: Instance-wise variable selection using neural networks," in *International Conference on Learning Representations*, 2019.

[9] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on

model interpretation," in *International conference on machine learning*. PMLR, 2018, pp. 883–892.

[10] P. Schwab and W. Karlen, "Cxplain: Causal explanations for model interpretation under uncertainty," *Advances in neural information processing systems*, vol. 32, 2019.

[11] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[12] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=S1jE5L5gl

[13] F. Liang, Q. Li, and L. Zhou, "Bayesian neural networks for selection of drug sensitive genes," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 955–972, 2018.

[14] O. Zaidan and J. Eisner, "Modeling annotators: A generative approach to learning from annotator rationales," in *Proceedings of the 2008 conference on Empirical methods in natural language processing*, 2008, pp. 31–40.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.