Privacy-Preserving Argumentative Explanations (Student Abstract)

Ungsik Kim*, Minjae Lee, Jiho Bae, Minje Kim, Sang-Min Choi, Suwon Lee

Gyeongsang National University, Jinju-si, Republic of Korea
{blpeng, wjdchs0129, dream_cacao_jh, alswp6597, jerassi, leesuwon}@gnu.ac.kr

Abstract

We propose a framework for privacy-preserving argumentative explanations using homomorphic encryption. This method applies the Cheon-Kim-Kim-Song scheme, along with a soft k-means adapted for encrypted computation, to generate explanations without exposing sensitive data. By leveraging GPU acceleration, speedups of approximately 470–670 times were achieved compared with CPU execution. Experimental results show that explanation fidelity is maintained for small- to medium-scale models, whereas significant degradation occurs in larger models. These findings suggest that our study provides an initial step toward enabling secure and trustworthy argumentative explanations under encryption while also highlighting the challenges that remain for generalizability to more complex models.

Introduction

As machine learning is being increasingly applied in high-stakes areas such as healthcare and finance, explaining the model outputs has become critical. Traditional explainable artificial intelligence (XAI) methods such as local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) and Shapley additive explanations (SHAP) (Lundberg and Lee 2017) indicate the importance of each feature; however, this is insufficient to convince the user of the reasonableness of the results.

Several argumentative explanations (Potyka 2021) that go beyond merely showing which features are important have been proposed to overcome this limitation, such as providing a structured form of arguments that include both grounds for a claim and counterarguments to it. This style of explanation explicitly presents both supporting grounds and counterarguments for a model decision, making the reasoning auditable in high-stakes settings such as credit or healthcare, rather than only highlighting which features are important.

However, argumentative explanations can cause problems in sensitive domains. The grounds and counterarguments of arguments often directly contain personal data, whereby the very process of generating explanations may lead to the disclosure of private information. Therefore, achieving explainability and privacy protection is a central challenge.

A related approach is federated learning, in which a model is trained on local devices and only aggregated model updates are shared with a central server, reducing direct exposure of raw data. By contrast, homomorphic encryption (HE) permits direct computation on encrypted data, enabling explanation generation without revealing the original inputs.

However, HE has inherent limitations: computations on ciphertext are dramatically slower than on plaintext, and common steps in explanation pipelines (e.g., ReLU or hard cluster assignment) are not directly compatible with CKKS. As a result, privacy-preserving argumentative explanation is often viewed as impractical beyond very small models. In this work, we build an end-to-end encrypted pipeline that replaces these steps with low-degree polynomial activations and a soft k-means variant, and we empirically measure both its runtime (CPU vs. GPU) and its faithfulness across model sizes.

The main contributions are as follows: (i) we provide an end-to-end pipeline that generates argumentative explanations entirely under CKKS without decrypting sensitive inputs; (ii) we adapt non-HE-friendly steps (ReLU, hard argmin in clustering) into low-degree polynomial activations and a soft k-means that runs directly on ciphertext; (iii) we quantify the practical feasibility of encrypted argumentative explanation by reporting both runtime (CPU vs. GPU, with 470–670$\times$ speedups) and faithfulness degradation as model size grows.

Methodologies

The model is first trained in plaintext as a standard multilayer perceptron (MLP) with ReLU activation, performing inference and explanation in the encrypted domain using the CKKS scheme. In this setting, the ReLU function is replaced by a low-degree polynomial $p(x) = x^2 + x$ to accommodate the constraints of the encrypted environment. Linear layers are computed with fixed weights using ciphertext–plaintext multiplications, additions, and rotation–accumulation patterns. For the clustering required in argumentative explanations, soft probability assignments are employed instead of hard assignments, implementing normalization and the reciprocals using Newton's approximation, for the execution

*Postal Address: Gyeongsang National University, 501 Jinju-daero, Jinju-si, Gyeongsangnam-do, 52828, Republic of Korea; Phone: +82-10-6444-2476

Model	Setting	Infer	Expl
MLP-S	Baseline	0.00 s	0.02 s
	Ours (GPU)	5.46 s	24.78 s
	Ours (CPU)	49.6 s	7613.56 s
MLP-M	Baseline	0.01 s	0.08 s
	Ours (GPU)	9.23 s	113.36 s
	Ours (CPU)	159.71 s	54,213.54 s
MLP-L	Baseline	0.01 s	0.38 s
	Ours (GPU)	20.67 s	645.19 s
	Ours (CPU)	1,451.96 s	438,213.15 s

Table 1: Comparison of inference and explanation generation speed across model sizes; Infer and Expl denote model inference speed and explanation generation speed, respectively. Under Setting, Baseline indicates inference and explanation generation in plaintext; Ours (GPU) indicates inference and explanation generation in ciphertext with GPU acceleration, whereas Ours (CPU) reports the time required when performing the computation on CPU.

to proceed without decryption. The multiplication and rotation operations, the main bottlenecks of CKKS, are executed in parallel on a GPU to reduce the execution time compared with CPU-only computations (Badawi et al. 2022). After clustering, weight aggregation follows the (Ayoobi, Potyka, and Toni 2023) method.

Experiments

We evaluate on a Breast Cancer tabular classification task. Experiments ran on an Intel Core Ultra 9, an RTX 5090, and 64 GB RAM, with $n = 2^{12}$ CKKS slots. We consider three MLPs: MLP-S (3 layers, 30 nodes each), MLP-M (5 layers, 50 nodes each), and MLP-L (7 layers, 100 nodes each). Inference and explanation were tested on 16 held-out samples.

Table 1 presents the inference and explanation generation speeds for both plaintext and ciphertext. In MLP-S, the propagation generation speed was approximately 300× faster than that on CPU; in MLP-M, approximately 470× faster; and in MLP-L, approximately 670× faster. Table 2 reports the evaluation results of the explanation quality for both plaintext and ciphertext. We employed two metrics: input–output unfaithfulness, which measured the degradation of the model outputs under explanation, and structural unfaithfulness quantified the degree of distortion in internal representation. As classical structural unfaithfulness relied on hard assignments, we extended it to the soft k-means setting by reconstructing activations from probabilistic memberships. Notably, explanation fidelity remained relatively close to that of the plaintext up to MLP-M; however, MLP-L exhibited severe degradation.

Future Work

In future work, we aim to reduce the fidelity drop observed in larger models such as MLP-L by exploring HE-friendly clustering mechanisms beyond soft k-means. We also plan to further optimize encrypted linear-algebra kernels on GPU-class hardware, with the goal of pushing encrypted argumentative explanation closer to interactive latency while preserving faithfulness.

Model	Setting	IO (↓)	Structural (↓)
MLP-S	Baseline	0.0108	0.1024
	Ours	0.0387	0.1386
MLP-M	Baseline	0.0098	0.1084
	Ours	0.0819	0.1402
MLP-L	Baseline	0.0499	0.4300
	Ours	0.4454	0.5172

Table 2: Comparison of unfaithfulness in plaintext and ciphertext reasoning across model sizes; IO denotes input– output unfaithfulness, and Structural denotes structural unfaithfulness. Baseline represents explanation quality in plaintext, and Ours represents explanation quality in ciphertext.

Conclusion

This study presents a framework for generating privacy-preserving argumentative explanations based on CKKS homomorphic encryption. By combining a probabilistic soft k-means with GPU-accelerated encrypted computation, we have demonstrated that structured explanations can be obtained without exposing sensitive data. Although our results show promising speed improvements over CPU computations, the explanation fidelity is degraded for larger models, indicating the need for further research. Overall, our work provides a proof-of-concept foundation for encrypted argumentative explanations and indicates future directions for improving scalability and reliability in sensitive domains such as healthcare and finance.

References

Ayoobi, H.; Potyka, N.; and Toni, F. 2023. SpArX: Sparse Argumentative Explanations for Neural Networks [Technical Report]. *arXiv preprint arXiv:2301.09559*.

Badawi, A. A.; Alexandru, A.; Bates, J.; Bergamaschi, F.; Cousins, D. B.; Erabelli, S.; Genise, N.; Halevi, S.; Hunt, H.; Kim, A.; Lee, Y.; Liu, Z.; Micciancio, D.; Pascoe, C.; Polyakov, Y.; Quah, I.; R.V., S.; Rohloff, K.; Saylor, J.; Suponitsky, D.; Triplett, M.; Vaikuntanathan, V.; and Zucca, V. 2022. OpenFHE: Open-Source Fully Homomorphic Encryption Library. Cryptology ePrint Archive, Paper 2022/915. https://eprint.iacr.org/2022/915.

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

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6463–6470.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " 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*, 1135–1144.