POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Privacy Preserving Object Detection

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Salvatore Buono

Advisor: Prof. Manuel Roveri

Co-advisors: Luca Colombo , Alessandro Falcetta

Academic year: 2023-2024

1. Introduction

With the rapid growth of cloud computing, the need for privacy-preserving techniques has become more critical than ever. While the cloud offers unmatched computational power and scalability, it also poses significant privacy risks. Data stored and processed in the cloud is vulnerable to unauthorized access, data breaches, and even malicious attacks. This is where homomorphic encryption (HE) comes into play, providing a solution that allows computations to be performed directly on encrypted data, meaning sensitive information never needs to be decrypted, even during processing. This is especially true in fields like satellite images analysis such as environmental monitoring, disaster management, and military surveillance, where personal and confidential data is handled. This thesis delves into the integration of privacy-preserving techniques within the context of deep learning, focusing particularly on adapting neural networks for object detection tasks using HE. While privacy-preserving methods have been extensively explored in image classification [3], their application in object detection has not been thoroughly investigated.

2. Research Objectives

The core objective of this thesis is to create and implement versions of object detection models that can function securely under HE, a type of encryption that allows computations to be performed on encrypted data without needing to decrypt it first. Achieving this involves several key goals. Firstly, the research focuses on adapting existing object detection architectures to be compatible with the HE scheme. Most used HE scheme are BFV and CKKS, with the last one which is more suitable for deep learning tasks since it is ideal for applications that tolerate approximations, whereas BFV is better suited for exact arithmetic on encrypted integers. This adaptation is crucial because it ensures that the models can process encrypted data without exposing sensitive information. Secondly, the thesis aims to optimize these models to reduce the computational overhead, which is a common challenge with HE due to its intensive resource requirements. By optimizing the models, the research demonstrate that it is possible to conduct privacy-preserving object detection efficiently with reduced inference and training times, while still ensuring that the security and privacy of the data are not compromised. Ultimately, the research strives to balance the

competing demands of maintaining privacy, ensuring accuracy, and optimizing computational efficiency.

3. State of the Art

The state-of-the-art techniques in privacy-preserving object detection face several significant limitations. First, computational overhead is a major challenge, particularly in methods such as Channel-Wise Homomorphic Encryption (CHE) [10] and BFV-based privacy-preserving CNNs [3]. While these approaches enable secure processing of data, they significantly slow down operations on encrypted datasets, making real-time processing and scalability difficult to achieve. A key drawback of the CHE method is its lack of batch processing capabilities, which limits throughput, especially for large-scale tasks like object detection. This limitation is problematic in environments where high-speed, real-time processing is critical. In addition, while lightweight models such as TinyissimoYOLO [7] offer computational efficiency, they often do so at the expense of accuracy. These models struggle to detect smaller or more complex objects, which can be a crucial requirement in satellite image analysis and environmental monitoring. Moreover, the Secure YOLOv3-SPP framework [11], which employs secret sharing for privacy preservation, introduces significant communication overhead, particularly in settings with limited bandwidth or network latency. This limits its applicability in real-time systems such as autonomous vehicles. Finally, BFV-based CNN approaches require complex parameter tuning like noise budget or polynomial degree and it can lead to degraded performance, making these methods challenging to implement in high-efficiency applications. Overall, while these privacy-preserving techniques advance the field, their current limitations restrict their practicality in many real-world scenarios where both efficiency and accuracy are required.

4. Methodology

The research methodology is structured and comprehensive, involving several critical steps to adapt and optimize object detection models for effective use in HE environments.

4.1. Network selection and analysis

Initially, two prominent object detection models, Faster R-CNN [9] and YOLO [8] were considered. However, after careful analysis, Faster R-CNN was not chosen as the main model for this research. The primary reason for this decision is that Faster R-CNN, while highly accurate, relies on a multi-stage process that introduces significant computational overhead. Faster R-CNN's architecture includes a region proposal network (RPN) that generates multiple candidate regions (bounding boxes) in the image, followed by a refinement step to classify the objects and predict their locations. This multi-stage approach, while precise, is computationally intensive and involves complex non-linear operations that are difficult to implement efficiently under HE.
Instead, this research opts for the YOLO (You Only Look Once) model [8], which processes the entire image in a single pass. YOLO is known for its speed and efficiency, making it a more suitable choice for cloud environments where both speed and resource efficiency are critical. Unlike Faster R-CNN, YOLO performs object detection and classification in one step, significantly reducing the computational load. This single-stage approach aligns better with the requirements of HE, as it minimizes the number of complex operations that need to be performed on encrypted data, thus improving performance and efficiency in the cloud.

4.2. Substitution of Non Linear Operation

One of the main challenges in adapting these models for HE is dealing with non-linear operations like ReLU activation and max-pooling. These operations are widely used in neural networks because they help the models learn complex patterns in data. However, they are not directly compatible with HE, which typically only supports operations which can be approximated with polynomials. To address this, the research replaces these non-linear operations with HE-compliant alternatives. For instance, ReLU activation can be approximated using polynomial functions [2], (Figure 1) and max-pooling can be substituted with average pooling [3]. These substitutions are essential to ensure that the models can still function effectively while processing encrypted data.

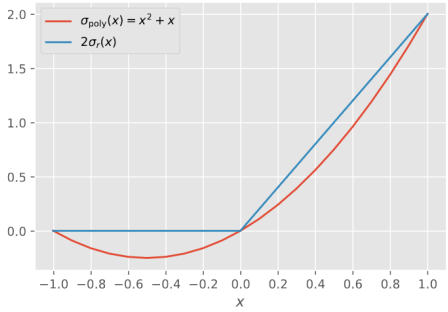

Figure 1: Approximation of relu with second order polynomial

4.3. Optimization of Network Complexity

In addition to modifying specific operations, the overall complexity of the network also needs to be optimized. HE operations are computationally expensive, meaning they require significant processing power and time [3]. Therefore, it is crucial to minimize the multiplicative depth of the network, i.e., the number of consecutive multiplications it performs, before the noise in the ciphertext grows too large, causing the decryption to fail. The research involves simplifying the network by reducing the number of layers and optimizing the operations to keep the multiplicative depth as low as possible. This step is vital to maintaining the model's performance while ensuring that it can operate efficiently within the constraints of HE.

4.4. Redesigning of Loss Function

Loss functions are used in training neural networks to measure how well the model's predictions match the actual results. In object detection, these functions often involve complex calculations, such as determining the overlap between predicted bounding boxes and the actual object locations (a metric known as Intersection over Union, or IoU). However, these calculations are not feasible in an HE environment due to their computational intensity. To overcome this challenge, the thesis proposes a simplified loss function that focuses on predicting the center points of objects rather than their full bounding boxes [5].

$$
\begin{aligned}
&\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&\quad + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
&\quad + \sum_{i=0}^{S^2} \mathbb{I}_{i}^{\text{obj}} \sum_{c \in \text{ classes}} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
$$

4.5. Ensemble and NMS optimization

To further enhance the performance of the HE object detection model, an ensemble method is employed. Ensemble learning involves combining multiple models to improve the overall accuracy and robustness of the predictions. In this thesis, different versions of the YOLO model, trained on various dataset initializations, are used to form an ensemble. Each model is independently trained, and their predictions are aggregated using weighted averaging. The weights assigned to each model's prediction are determined by the model's validation performance. This ensures that models with better accuracy contribute more to the final prediction.
Additionally, Non-Maximum Suppression (NMS) is adapted to work with the ensemble predictions. While original NMS reason in terms of overlapping bounding boxes, this version is lighter since it focuses only on distances between centers and deletes ones which are too close since they can represent the same object. This combination of ensemble learning and NMS significantly enhances the model's ability to detect objects accurately, particularly in scenarios with complex or overlapping objects.

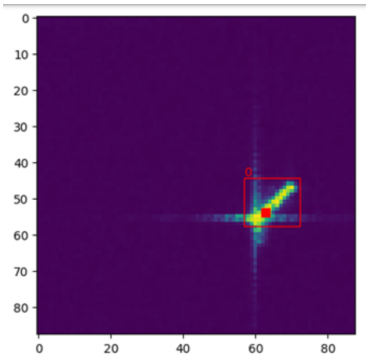

Figure 2: Accuracy of the network w.r.t target bounding boxes

The use of ensemble methods helps to improve the detection accuracy of the HE-adapted YOLO model while maintaining its computational efficiency, making it even more suitable for deployment in cloud environments.

4.6. Integration with HE

After making the necessary modifications and optimizations, the adapted networks are integrated within a HE framework using the Pyhelayers library [1]. The workflow begins with the initialization step, where the CKKS scheme is selected for its ability to handle approximate arithmetic operations on encrypted data, particularly useful for neural network operations. The system is optimized based on batch size, determining how many samples are processed simultaneously during inference. Next, the test images are encrypted using Pyhelayers, which simplifies HE tasks for users. Once encrypted, the model performs predictions on an untrusted server that cannot access the original data, model parameters, or results, as it lacks the secret key. To improve accuracy, predictions from different models can be combined for an ensemble result. After the encrypted inference, results are sent back to the trusted client, decrypted, and decoded. The decrypted information includes key data, such as the center coordinates of detected objects relative to the image dimensions, as used in models like YOLO [8]. By combining encrypted computation, ensemble methods, and NMS, this technique is particularly well-suited for cloud-based applications where data privacy is paramount.

5. Results

The experimental results provide strong evidence of the efficiency and accuracy of the privacy-preserving object detection method proposed in this thesis. A series of experiments were conducted to evaluate the performance of the adapted YOLO model, integrated with HE under the CKKS scheme, in a cloud-based environment. The results focus on three key aspects: privacy preservation, computational efficiency, and object detection accuracy.

5.1. Privacy Preservation in Cloud Computing

The primary goal of maintaining data privacy throughout the computational process was achieved. The experiments demonstrated that are not decrypted during the object detection process on the server side. The entire inference process was conducted on encrypted data, and the encrypted results were successfully sent back for decryption on the client side. This ensures that the privacy of the images and their labels was fully preserved. Graphs such as Figure 3 in the thesis illustrate the workflow of data encryption [6], computation in the cloud, and decryption on the client side. These visualizations provide a clear representation of how privacy is maintained throughout the process.

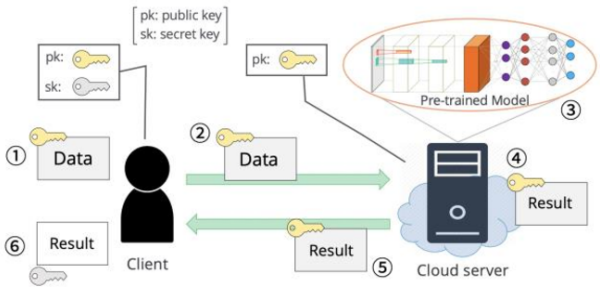

Figure 3: HE workflow

5.2. Computational Efficiency

One of the primary challenges when using HE is the added computational overhead. To assess the model's efficiency, several experiments were conducted to measure inference time, both with and without HE. Table 1 highlights how the streamline and optimization impacts on perfomances changing from a YOLO [8] to FOMO [5] where only the center coordinates are predicted. As we can see, when we want to compute the size

of bounding box, there is an increase in inference time due to the encrypted operations, but the optimizations made to the model—such as reducing the multiplicative depth and replacing non-linear operations—led to significant performance improvements. The table shows that, despite the encryption overhead, the FOMO model is able to perform object detection in feasible time frames for cloud-based applications. This performance makes it practical for deployment in real-world scenarios where privacy is a critical concern.

Model	Plain	Enc	RAM
TinyYOLO	459ms	2h30m	120GB
TinyissimoFOMO	74ms	12m	30GB

Table 1: Computational complexity of different models

5.3. Object Detection Accuracy

A key focus of this thesis is the adaptation of the TinyissimoYOLO architecture to HE environments, particularly by simplifying the object detection task to predicting object centers rather than full bounding boxes. This simplification reduces computational complexity, which is crucial for maintaining feasible processing times in HE contexts [4]. These experiments evaluate the impact of various combinations of activation functions (ReLU, Square, Linear) and pooling strategies (MaxPooling, AvgPooling) on both bounding box and center prediction tasks. The mAP (mean Average Precision) metric, which considers a true positive (TP) when the predicted center lies within the target bounding box, is used to compare the models. The results show a clear distinction between how different architectures perform in HE environments and how activation functions impact on performances. Building the best polynomial approximation of non-linear function (in the table 2 called *linear*) can lead to better gradient propagation and it allows models to capture complex relationships. The experiments demonstrate that simplifying the object detection task, using a fixed backbone architecture, significantly improves performance in HE environments.

	Activation	Pooling	mAP
Bbox	ReLu	MaxPooling	0.3
Bbox	Square	AvgPooling	0.005
Bbox	Linear	AvgPooling	0.09
Center	ReLu	MaxPooling	0.67
Center	Square	AvgPooling	0.29
Center	Linear	AvgPooling	0.586

Table 2: YOLO and FOMO comparison w.r.t HE compatible operator.

5.4. Impact of Ensemble Method on Accuracy

The ensemble method proved particularly effective in enhancing detection accuracy, as shown in Figure 4. This experiment highlights the accuracy improvement achieved by combining the predictions of multiple models as the average distance between the predicted and target boxes is smaller compared to individual models. The ensemble method contributed to more robust predictions, particularly in complex object detection scenarios, by reducing the variance of individual models and improving overall detection reliability. This is especially useful when dealing with encrypted data, where small inaccuracies can sometimes lead to larger prediction errors.

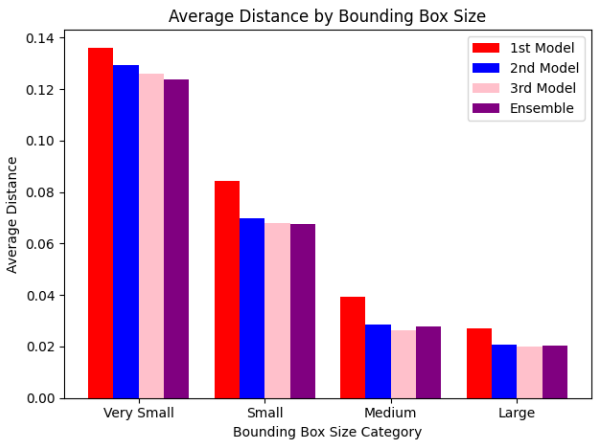

Figure 4: Ensemble performances with respect to box categories

6. Conclusions

This thesis makes significant strides in the field of privacy-preserving deep learning by providing

practical solutions for secure object detection using HE. The research shows that it is possible to achieve efficient and secure object detection without sacrificing the accuracy of the models. By optimizing and adapting well-established object detection models, this thesis opens the door to further exploration in this area, particularly in optimizing other neural network architectures for HE environments. The findings of this research have broad implications, offering a pathway to extending these techniques to other domains where privacy is of primary importance.

7. Acknowledgements

I would like to express my deepest gratitude to everyone who supported me throughout these incredible academic years. First and foremost, I am profoundly grateful to my family, who not only gave me the opportunity to pursue my studies in a different city but also stood by me with huge support whenever I needed it. A special thank you goes to my girlfriend and closest friends, whose presence brought joy and comfort during the most challenging times, and whose encouragement constantly inspires me to become a better person. Finally, I extend my sincere thanks to everyone who contributed to this research. To my professor and his assistants, I am especially thankful for their invaluable guidance, insightful suggestions, and for fostering my critical thinking throughout this journey. Your mentorship has been instrumental in shaping this work.

References

[1] Ehud Aharoni, Allon Adir, Moran Baruch, Nir Drucker, Gilad Ezov, Ariel Farkash, Lev Greenberg, Ramy Masalha, Guy Moshkowich, Dov Murik, Hayim Shaul, and Omri Soceanu. Helayers: A tile tensors framework for large neural networks on encrypted data. *Proceedings on Privacy Enhancing Technologies*, 2023(1):325–342, January 2023.

[2] Ramy E. Ali, Jinhyun So, and A. Salman Avestimehr. On polynomial approximations for privacy-preserving and verifiable relu networks, 2024.

[3] Alessandro Falcetta and Manuel Roveri. Privacy-preserving deep learning with homomorphic encryption: An introduction. *IEEE Computational Intelligence Magazine*, 17(3):14–23, 2022.

[4] Shruthi Gorantala, Rob Springer, and Bryant Gipson. Unlocking the potential of fully homomorphic encryption. *Commun. ACM*, 66(5):72–81, apr 2023.

[5] Peter Ing. Introducing faster objects more objects aka fomo, 03/29/2022. Accessed: 08/29/24.

[6] Takumi Ishiyama, Takuya Suzuki, and Hayato Yamana. Highly accurate cnn inference using approximate activation functions over homomorphic encryption. *2020 IEEE International Conference on Big Data (Big Data)*, pages 3989–3995, 2020.

[7] Julian Moosmann, Marco Giordano, Christian Vogt, and Michele Magno. Tinyissimoyolo: A quantized, low-memory footprint, tinyml object detection network for low power microcontrollers. In *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, June 2023.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[10] Tianying Xie, Hayato Yamana, and Tatsuya Mori. Che: Channel-wise homomorphic encryption for ciphertext inference in convolutional neural network. *IEEE Access*, 10:107446–107458, 2022.

[11] Yongjie Zhou, Jinbo Xiong, Renwan Bi, and Youliang Tian. Secure yolov3-spp: Edge-cooperative privacy-preserving object detection for connected autonomous vehicles. *2022 International Conference on Networking and Network Applications (NaNA)*, pages 82–89, 2022.