Lightweight and Efficient Privacy-Preserving Multimodal Representation Inference via Fully Homomorphic Encryption

Zhaojue Li[1], Yingpeng Sang[1](✉), Xinru Deng[1], and Hui Tian[2]

[1] School of Computer Science and Engineering, Sun Yat-sen University,
Guangzhou, China
{lizhj33,dengxr3}@mail2.sysu.edu.cn, sangyp@mail.sysu.edu.cn

[2] School of Information and Communication Technology, Griffith University,
Brisbane, Australia
hui.tian@griffith.edu.au

Abstract. Machine learning models are now being widely deployed in clouds, but serious data leakage problems are also exposed when dealing with sensitive data. Homomorphic encryption (HE) has been used in the secure inference on unimodal private data because of its ability to calculate encrypted data. Although the privacy protection of multimodal data is of great significance, there is still no privacy-preserving inference scheme for multimodal data. In this work, we propose a lightweight and efficient homomorphic-encryption based framework that enables privacy-preserving multimodal representation inference. Firstly, we propose an HE scheme based on Tensor Fusion Network, which can perform encrypted multimodal feature fusion. Then we propose a pre-expansion method and a packaging method for multimodal data, which can effectively reduce the time delay and data traffic of homomorphic computing. The experimental results show that our encryption inference method has almost no loss of accuracy and obtains an F1 score of 0.7697, while using less than 220KB of data throughput and about 0.91 s of evaluation time.

Keywords: Fully Homomorphic Encryption · Privacy-Preserving Machine Learning · Multimodal Privacy

1 Introduction

Our perception of the environment is inherently multimodal, involving multiple sensory channels such as vision, hearing and taste. A modality refers to the way in which something happens or is experienced, and research is considered multimodal when it involves more than one of these channels. Machine learning approaches based on multimodal data have emerged as a rapidly growing field of research [12,26,27,29]. Substantial research [4,21] has shown that classifiers

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
N. T. Nguyen et al. (Eds.): ACIIDS 2023, LNAI 13995, pp. 307–321, 2023.
https://doi.org/10.1007/978-981-99-5834-4_25

based on multimodal data outperform those based on unimodal data, which is consistent with the way in which humans perceive and comprehend the world.

Transmitting multimodal sensitive data from client to server typically entails privacy risks, as the server may be malevolent or susceptible to attacks by third parties. Moreover, because multimodal data contains rich cross-information, the privacy damage caused by leakage is more severe than that caused by unimodal data. Thus, there is a significant motivation to design a privacy-preserving strategy for multimodal machine learning that can ensure the privacy of clients who use cloud computing services. A highly relevant technology is Fully Homomorphic Encryption (FHE) [15]. FHE is an encryption system that enables data owners to encrypt their data and authorize third parties to perform calculations on it. Third parties can perform several calculations but are not authorized to access the original data, thus preserving the privacy of consumers. However, practical applications require specialized solutions to reduce the high computing and transmission costs associated with FHE. Previous works have provided specialized homomorphic schemes for unimodal machine learning [2,8,11].

In this paper, we propose, for the first time, a privacy-preserving prediction approach for multimodal representation based on the FHE scheme CKKS [10] and Tensor Fusion Network (TFN) [32]. In this method, the server accepts the encrypted multimodal representation sent by the client, performs fusion and model evaluation, and feeds back the evaluation results to the client. Only the client can decrypt the results and obtain the inference. The main contributions of our study are as follows:

- This research is the first to propose a multimodal representation inference approach based on FHE. Specifically, we implement the multimodal feature fusion network in the ciphertext state.
- We provide a pre-expansion method to reduce the computational complexity of the homomorphic tensor feature fusion layer.
- We utilize the rotation operation of CKKS and the unoccupied slot in ciphertext, to pack data of multiple modes in the same ciphertext. This significantly decreases the amount of data transmission required. To optimize the ciphertext matrix multiplication with the highest latency, we also leverage multithreading to achieve a 3.5-times acceleration.

In the rest of the paper, we summarize the related work in Sect. 2, and present the background knowledge in Sect. 3. In Sect. 4 we present our proposed approach in details, which is followed by performance evaluation and experimental results in Sect. 5. Finally, we conclude the paper in Sect. 6.

2 Related Work

2.1 Multimodal Machine Learning

The employment of multimodal data provides human beings with a comprehensive and multifaceted understanding, thereby facilitating them to make more informed decisions [4,12]. Despite the convenience brought by cloud services, the privacy concern regarding multimodal data is increasingly urgent. To address this issue, Cai et al. [9] proposed implementing differential privacy in the representation after multimodal fusion to protect the privacy of the original training data. However, the introduction of noise by differential privacy inevitably compromises the utility of the data and the accuracy of the model inference. Moreover, it is noteworthy that this method cannot protect user privacy data during the inference process.

2.2 Homomorphic Encrypted Neural Network

With the increasing application of machine learning in education, finance, and other fields that deal with sensitive customer data, there is a growing need for privacy protection in machine learning algorithms that make accurate predictions. To address this issue, several cryptographic techniques have been proposed, such as Trusted Execution Environments [20], Secure Multi-Party Computation (SMPC) [28] and Homomorphic Encryption (HE) [15]. Each method has its own tradeoffs in terms of calculation, accuracy, and security. Among them, schemes based on Fully Homomorphic Encryption (FHE) can generally achieve quantum security, which is the most rigid security model [1].

FHE was first proposed by Rivest et al. [24], and Gentry [15] proposed the first generation of FHE systems. Despite allowing more homomorphic multiplication and addition operations, practical applications still require too much computation. To address this issue, several practical leveled homomorphic encryption schemes have been proposed, such as the integer-based BGV algorithm [7], BFV [14], and the complex scheme CKKS [10]. These schemes can perform homomorphic computation within the pre-set maximum multiplication depth. CKKS permits approximate HE operations for real numbers, making it a suitable choice for the inference task of machine learning with a fixed number of layers.

Dowlin et al. [16] proposed CryptoNets, which proved the feasibility of using HE for private neural network inference. However, CryptoNets have two limitations. The first is on the time cost, although it supports high-throughput prediction, the prediction of a single sample still takes 205 s. The second is on the width of the network. CryptoNet encodes each node of the network into a separate ciphertext, so it needs a lot of memory to support it. In order to solve these problems, LoLa [8] encrypted the entire network layer, significantly reducing memory requirements and achieving single sample prediction in 2.2 s. Furthermore, homomorphic schemes have been proposed for text classification [2] and audio similarity calculation [22] in addition to image classification. However, these schemes only consider the homomorphic implementation of unimodal machine learning. To the best of our knowledge, there is currently no homomorphic encryption scheme available for multimodal machine learning and multimodal feature fusion.

2.3 Homomorphic Representation Inference

The aforementioned work can be used to encrypt shallow networks through homomorphic inference. However, there are two primary limitations when Fully Homomorphic Encryption (FHE) is applied to deep network models: the growth of noise and the growth of ciphertext size. Each ciphertext contains noise that increases with each homomorphic operation. Therefore, too many operations increase the noise to the point where the decryption may not be correct. Secondly, the operation of the HE scheme can double the size of the encryption parameter without bootstrapping, resulting in a large ciphertext that increases memory requirements and causes greater latency. To address these issues, LoLa [8] proposed using deep representations for encrypted inference. Customers convert the original data into deep representations locally through the feature extraction network. Their prediction only requires a shallow network model, which is more suitable for low-latency homomorphic implementation. Chou et al. [11] proposed extracting deep representations from the screenshot of the original phishing website locally, encrypting them, and sending them to the cloud for homomorphic logistic regression calculation, achieving a low-delay and secure homomorphic inference scheme.

3 Preliminaries

3.1 Fully Homomorphic Encryption (FHE)

Fully Homomorphic Encryption (FHE) is a powerful encryption method that allows ciphertext computation with minimal loss of accuracy upon decryption. This capability makes it useful for secure computing outsourcing: the client encrypts the data and sends it to a third party for computation, where the third party cannot access the plaintext data. After receiving the encrypted output, the client decrypts it to obtain the computation result. Specifically, the encryption function is denoted by Enc, and plaintext data by x and y. Then, $Enc(x + y) = Enc(x) \oplus Enc(y)$ and $Enc(x * y) = Enc(x) \odot Enc(y)$, where $\oplus$ and $\odot$ represent homomorphic addition and multiplication, respectively.

While HE allows ciphertext computation to obtain the encrypted output with practically little loss of accuracy, it adds noise to plaintext data, and homomorphic operations increase noise continuously. Once the noise reaches a certain threshold, the correct plaintext result cannot be decrypted. Although a primitive bootstrapping method [15] can refresh noise, it is limited by the massive amount of computation required. A more practical technique is to use leveled homomorphic encryption [5], which allows multiple addition and multiplication operations at a predetermined maximum multiplication depth. Since the calculation times of neural networks are also determined, leveled homomorphic encryption has been widely used in encrypted machine learning inference tasks. In this paper, we choose the CKKS leveled homomorphic encryption method, which is specifically designed to handle real numbers and approximate calculations.

3.2 The Levelled FHE Scheme - CKKS

In the following paragraphs, we will briefly introduce the CKKS scheme. Let N be a power of two, and $R = \mathbb{Z}[X]/(X^N + 1)$ be the ring integer of the $2N$-th cyclotomic polynomial. For some small prime integers p_i, let $q_L = \prod_{i=1}^{L} p_i$ and $R_{q_L} = \mathbb{Z}_{q_L}[X]/(X^N + 1)$ consists of the polynomials whose coefficients are moduloq_L. Here, L represents the preset maximum multiplicative depth, and a message $m \in \mathbb{C}^{N/2}$ is encoded and encrypted into R_{q_L}, followed by homomorphic addition and multiplication. Each multiplication consumes a layer of depth and rescales the ciphertext of R_{q_l} into $R_{q_{L-1}}$. Another essential operation enabled by ciphertext is rotation, which allows encrypted elements to rotate in $N/2$ slots.

One advantage of CKKS is on its SIMD feature, that is, one single homomorphic addition (or multiplication) among ciphertexts can attain a corresponding addition (or element-wise product) of two vectors in plaintexts. Suppose that m is a k-dimensional vector. When $k < N/2$, m will be padded with zeros to size $N/2$. As the homomorphic operation operates bit by bit on all slots, the operation is highly inefficient when $k << N/2$. The optimization methods introduced later in this paper will fully utilize the free spaces in the ciphertext to improve computational efficiency without incurring additional space consumption.

3.3 Homomorphic Linear Layer

Here we introduce the implementation of the homomorphic linear layer [5]. The linear layer is one of the most important network layers in the machine learning model. A linear layer consists of a vector-matrix multiplication and an addition of a bias. Traditionally, each column of the matrix can be packaged and multiplied by the ciphertext vector, and the result can be summed bit by bit. This practice causes the output to be spread across multiple ciphertexts rather than stored under a single ciphertext. Halevi et al. [17] proposed a matrix multiplication that is realized by matrix diagonalization. Let $n = N/2$ denotes the number of slots in ciphertext c, matrix $M \in R^{n \times m}$. Firstly M is decomposed into n vectors in diagonal order, in which the j'th element in the i'th diagonal $diag[i][j] = M[(i+j) mod\, n][j]$. Then we have $M.c = \sum_{i=0}^{n-1} diag(i) \odot Rotate(c, i)$, where $\odot$ denotes the coefficient wise vector multiplication and $Rotate(c, i) = (c_i, c_{i+1}, \ldots, c_0, c_1, \ldots, c_{i-1})$ is the rotation of c by shifting i slots to the left. In practice, in order to get the correct rotation result, it is necessary to copy the encryption vector [19], then $c = (c_0, c_1, \ldots, c_{n-1}, c_0, c_1, \ldots, c_{n-1}, 0, \ldots, 0)$. The complexity of vector-matrix multiplication is $O(n)$, and the computational cost is the largest in the homomorphic linear layer. Since the nonlinear activation function (such as ReLu) cannot be calculated in the ciphertext state, the standard practice [3] is to substitute it with the square activation function. The last layer of the network model is a homomorphic dot product layer, which is first multiplied by a k-dimensional vector. All elements are then added to the first element of the output ciphertext by $log(k)$ rotations. It should be noted that the final activation function is computed locally after decryption by the client [2].

3.4 Multimodal Fusion Representation Learning

The original multimodal data contains a large amount of redundant information, and the feature vectors of each modality are initially located in different subspaces. This can impede the learning of data in subsequent models. To address this issue, representation learning has been proposed as a solution [6]. This technique maps input data to a low-dimensional representation, enabling efficient learning. Currently, unimodal representation learning is widely used in Natural language processing (NLP)[13] and Computer Vision (CV) [18]. In our work, we employ an appropriate encoder network for each modality of the data, and fuse the resulting low-dimensional feature vectors. The validity of the representation is ensured by the convergence of network parameters.

Multimodality Fusion Technology (MFT) [23] involves fusing the feature vectors of each modality to create a more effective representation for subsequent networks. One simple method [33] for feature fusion is direct concatenation, but it can be challenging to capture the interaction information and nonlinear relationships between different modalities. To address this issue, Zadeh et al. proposed Tensor Fusion Network [32]. This approach models each sub-modality feature as different dimensions of a Cartesian space. Therefore, the fusion process between different modes can be achieved through the tensor cross-product, as shown in Eq. (1):

$$z = \begin{bmatrix} v_1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} v_2 \\ 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} v_n \\ 1 \end{bmatrix} \tag{1}$$

where z represents the output after TFN, v_i denotes different modalities and $\otimes$ denotes the outer product operator. To ensure that the fusion representation contains both the cross-information from dual mode to n mode and the independent information of each unimodal feature, an element of 1 is spliced at the end of each modality representation.

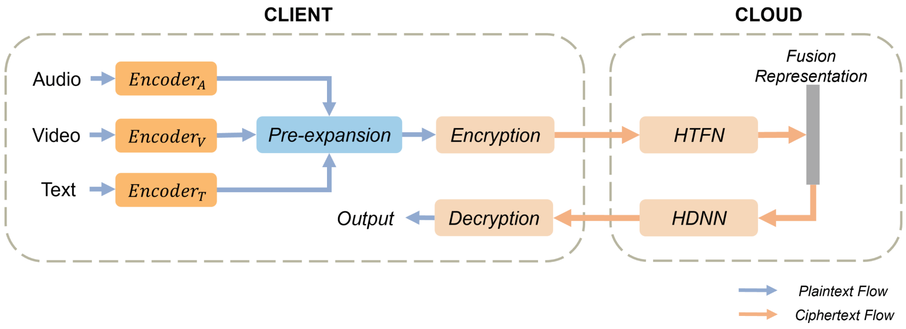

Fig. 1. HE-based Multimodal Representation Inference System

4 The Proposed Approach

4.1 System Model

This section introduces our system model and task settings. A client possesses multimodal raw data and uses the corresponding encoder for feature extraction. Subsequently, the client performs pre-expansion and transmits the ciphertext to the cloud. In the cloud, the encrypted fusion representation is obtained through the homomorphic tensor fusion network (HTFN). The homomorphic dense neural network (HDNN) performs the final evaluation, and the encryption result is delivered back to the client. The system model is illustrated in Fig. 1. Throughout the computing process, the server is limited to operating solely on the ciphertext and cannot access the sensitive data.

Consistent with current private reasoning tasks, such as those described in [8, 11, 22], the complete model network utilizes plaintext data during the common training phase and encrypted data during the inference phase. In this task, we implicitly assume that the client has the computational capacity to preprocess the raw data and execute the encryption/decryption tasks.

4.2 Homomorphic TFN

In TFN, the fusion features of a high-dimensional Cartesian product are obtained by multiplying the eigenvectors of each modality twice. However, in the ciphertext state, only the operation between vectors is supported, and the result of the Cartesian product cannot be retrieved directly. One possible solution, related to CryptoNets, is to encrypt each element of each feature independently and send it to the server for multiplication one by one. However, this approach is computationally expensive since the fusion requires computing hundreds of ciphertexts. An optimization for this method is *packing one modality* (POM) into a single ciphertext and multiply it with the elements of other features one by one. Even with this optimization, the computational complexity and the amount of output ciphertext are still dependent on the length of the feature.

To solve this issue, we propose a *pre-expansion* processing method. As the fusion features of the higher-dimensional Cartesian product will be flattened as input to the linear layer, and the SIMD feature of CKKS makes the bit-by-bit multiplication between ciphertexts efficient, we expand each representation to the fused length before encryption. In the simplest case of two modalities, pre-expansion will repeat each bit of one modality for L_1 times, given the other modality has a length of L_1. The other modality will also be expanded so as to be aligned with the previous modality, and finally both modality will be with a length of $L_1 \cdot L_2$, given L_2 is the length of the first modality. This will ensure that only one ciphertext multiplication is needed in the HTFN.

Let's consider the most widely used fusion of video, audio, and text, as an example. As shown in Fig. 2, the lengths of text, video and audio modality are 2, 2, 3, respectively. Firstly each bit of the video modality is repeated by 2 times, so that each bit can be aligned with the whole length of the text modality. Then correspondingly the lengths of video and text modalities will both be expanded to 4. In the same way, each bit of the audio modality is repeated by 4 times, and finally, all the three modalities will be expanded to be with a length of 12.

Let m denote the number of modalities, and L_i denote the length of the i'th modality feature. After pre-expansion, all modalities will be with a length of $\prod_{i=1}^{m} L_i$, and $m-1$ ciphertext multiplications are required to compute the HTFN. As shown in Table 1, the number of required ciphertext multiplications in HTFN can be reduced using pre-expansion, and the resulting fused representation can be stored in a single ciphertext. In Sect. 3.2, we mentioned that the CKKS ciphertext has $N/2$ slots, which is significantly larger than the length of each feature. Therefore, pre-expansion can take advantage of the vacant slots in the ciphertext without additional space requirements.

Table 1. Performance Comparison of HTFN Under Different Pretreatment Methods.

Pretreatment Method	Multiplications	Output Size
CryptoNets	$\prod_{i=1}^{m} L_i - 1$	$\prod_{i=1}^{m} L_i$
POM	$\prod_{i=1}^{m-1} L_i$	$\prod_{i=1}^{m-1} L_i$
Pre-expansion	$m-1$	1

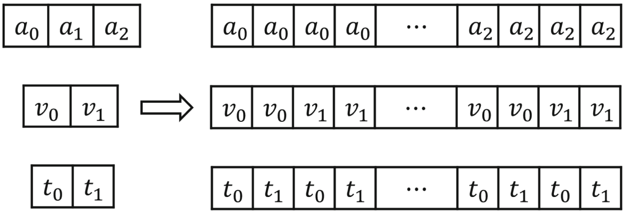
a_0 a_1 a_2
v_0 v_1
t_0 t_1
a_0 a_0 a_0 a_0 ⋯ a_2 a_2 a_2 a_2
v_0 v_0 v_1 v_1 ⋯ v_0 v_0 v_1 v_1
t_0 t_1 t_0 t_1 ⋯ t_0 t_1 t_0 t_1

Fig. 2. An Illustration of Pre-expansion on 3 Modalities

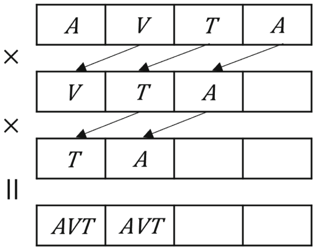

Fig. 3. An Illustration of HTFN with our Packing Method based on Ciphertext Rotations

4.3 Other Optimizations

The client sends a ciphertext for each modality to the cloud. As the number of modalities increases, the size of the ciphertexts also increases, resulting in a linear increase in encryption time and data transmission. To solve this problem, we concatenate all the feature vectors after the pre-expansion and finally encrypt them into a single ciphertext. In the ciphertext state, the ciphertext of each feature can be decoupled by two CKKS rotation operations. As mentioned in Sect. 3.3, the subsequent homomorphic linear layer computation requires a copy of the original vector. This can be achieved by concatenating the features of the first modality.

As shown in Fig. 3, assuming that the features fused by the three modalities have n dimensions, we perform two consecutive rotation operations to obtain two ciphertexts, where the first n bits of each ciphertext store the corresponding modality representation. Through this packaging method, the transmission of the original n ciphertexts can be optimized to a single ciphertext, which significantly decreases the client's encryption computation cost and transmission delay. Finally, to optimize the homomorphic linear layer with the longest computation time, we employ a multi-threaded approach for the n-times rotation and homomorphic multiplication. Algorithm 1 outlines the complete process of HTFN, where $\oplus$ and $\odot$ represent homomorphic addition and multiplication, respectively. For the sake of simplicity, we omit the relinearization and rescale operations that are mandatory after homomorphic multiplication and evaluation.

5 Experiments

5.1 Experimental Setup

We use two commonly used multimodal datasets for our experiments. Multimodal Corpus of Sentiment Intensity (CMU-MOSI) [33] contains 2199 movie reviews on YouTube video blog, which came from 89 narrators of ages among 20 to 30 years. Each has audio, text, and video information available. The videos were labeled $[-3, 3]$ with seven categories ranging from negative to positive affective tendencies. Different from the original CMU-MOSI dataset, we use BERT preprocessing to obtain more accurate text information [31]. CH-SIMS [30] is a Chinese multimodal emotion analysis data set. It cuts out 2281 video clips from different movies and television works and collates the data in audio, text and video modalities. It contains 474 different speakers with a wide range of characters and ages. Each video clip has five categories of emotional intensity values between $[-1, 1]$.

Algorithm 1: Homomorphic Multimodal Representation Inference

```
Input: ciphertext $c$ encrypting vector $v \in R^{m.k}$, where $m$ is the number of
       modalities, linear layer weight $W \in R^{k \times z}, y \in R^z$, bias $b_1 \in R^z, b_2 \in R$
Output: ciphertext $s$ encrypting the evaluation result
Step 1: Homomorphic Tensor Fusion Layer
$r \leftarrow c \in R^N$;
for $i = 1$ to $m - 1$ do
    $r \leftarrow r \odot Rotate(c, k * i)$;
end
Step 2: Homomorphic Linear Layer
$s \leftarrow 0 \in R^N$
for $i = 0$ to $k - 1$ do
    for $j = 0$ to $z - 1$ do
        $D[i][j] \leftarrow W[(i + j) mod\ k][j]$;
    end
    $s \leftarrow s \oplus D[i] \odot Rotate(r, i)$;
end
$s \leftarrow s \oplus b_1$;
$s \leftarrow s \odot s$;
Step 3: Homomorphic Dot Product Layer
$s \leftarrow s \odot y$;
for $i = \lceil log(z) \rceil$ to 1 do
    $t \leftarrow Rotate(s, 2^i)$;
    $s \leftarrow s \oplus t$;
end
$s \leftarrow s \oplus b_2$;
return s
```

For network settings, as in the previous work of [32], we choose LSTM to extract text features. DNN with three hidden layers is used for audio and video features. Following this, we use two hidden layers to make the final prediction of the fusion representation. It is worth noting that the maximum fusion representation length should not exceed 2048. Considering the actual usability, our experiment uses a smaller length, 512. The setting of other hyperparameters and the partitioning of data sets are consistent with the TFN experimental settings in MMSA [31]. After the model training, we use the SEAL library [25] to carry out the inference task of homomorphic encryption.

5.2 Experimental Results

In Table 2, we compare the inference results of the model before and after encryption on the test set using the F1 score and MAE as evaluation metrics, which are commonly used in multimodal machine learning. Results demonstrate that our encryption inference method is highly effective, as our model exhibits no performance loss compared to plaintext inference up to four decimal places of precision.

Table 3 illustrates the enhancements made by the proposed optimization method concerning data throughput and inference time in the cloud. We began with the idea of CryptoNets, which encrypts each element individually. Although it supports batch prediction of multiple samples, the time for a single calculation is approximately 150 s. Lola [8] is a low-delay computing model proposed by CryptoNets for single-sample calculations, but it does not support multimodal feature fusion. In POM, each element of the two features (lengths of 4 and 16) had to be packed into a separate ciphertext since TFN necessitates element-wise multiplication. The computation time has been reduced to 4.43 s, but for scenarios with high feature lengths, this time will increase rapidly and require larger data transfers. With pre-expansion, feature fusion can be completed with only two ciphertext vector multiplications, enabling the use of smaller encryption parameters like $N = 8192$ for further optimization while ensuring security. Furthermore, packing the features of the three modalities into a single ciphertext reduces the amount of data transmission to 1/3, which must be decoupled by two rotation operations on the server. Multithreading can effectively reduce the delay of ciphertext vector-matrix multiplication for the first homomorphic linear layer after feature fusion. Finally, the proposed optimization method yields an inference time of approximately 0.91 s for a single sample and the data transfer volume is 211.88 KB.

We present a detailed comparison of the performance of HTFN implemented under three preprocessing methods in Table 4. The number of multiplications required by our approach is linearly dependent on the number of modes, whereas the other two methods show a rapid increase in the number of multiplications as the feature length increases. Our approach, combined with two rotations and packaging, compactly stores both input and output in a single ciphertext, thus reducing the encryption burden on the client and the computational cost on the server. Experimental results demonstrate that our method can maintain stable and efficient computing performance in practical multimodal scenarios.

Table 2. Encrypted Test Set Inference Results for CMU-MOSI and SIMS.

Dataset	F1 score	MAE	Accuracy Loss
CMU-MOSI	0.7467	1.0400	No
SIMS	0.7697	0.4356	No

Table 3. Model Performance - Data Throughput and Timing.

Method	Data Throughput	Computation Time
CKKS	6.03 MB	149.21 s
CKKS + POM	2.70 MB	4.43 s
CKKS + pre-expansion	637.14 KB	3.12 s
CKKS + pre-expansion + packed	211.88 KB	3.14 s
CKKS + pre-expansion + packed + multi-thread	211.88 KB	0.91 s

Table 4. Performance Comparisons of HTFN.

	Ours	POM	CryptoNets
Multiplication	2	68	544
Rotation	2	-	-
Input Size	427.08 KB	8.77 MB	12.09 MB
Output Size	282.95 KB	17.92 MB	141.56 MB
Computation Time	0.036 s	1.051 s	8.407 s
Encryption Time	0.006 s	0.212 s	0.350 s

6 Conclusion

In this paper, we propose the first multimodal representation inference method based on Fully Homomorphic Encryption (FHE). Our method provides complete protection of multimodal data privacy and enables private prediction tasks on the cloud server. We propose a pre-expansion method and a homomorphic TFN scheme using only two ciphertext multiplications. We also propose a variety of optimization schemes that not only improve the computational efficiency of ciphertext but also reduce the amount of data transmission. The experimental results show that our method has the advantages of low computation and communication costs. In the future, we will continue to improve the performance of the model and extend it to multi-user application scenarios using multi-key FHE.

Acknowledgement. This work was supported by the Key-Area Research and Development Program of Guangdong Province (No. 2020B010164003), China.

References

1. Albrecht, M.R., et al.: Homomorphic encryption standard. IACR Cryptol. ePrint Arch., 939 (2019). https://eprint.iacr.org/2019/939
2. Badawi, A.A., Hoang, L., Mun, C.F., Laine, K., Aung, K.M.M.: PrivFT: private and fast text classification with homomorphic encryption. IEEE Access **8**, 226544–226556 (2020). https://doi.org/10.1109/ACCESS.2020.3045465
3. Badawi, A.A., et al.: Towards the AlexNet moment for homomorphic encryption: HCNN, the first homomorphic CNN on encrypted data with GPUs. IEEE Trans. Emerg. Top. Comput. **9**(3), 1330–1343 (2021). https://doi.org/10.1109/TETC.2020.3014636
4. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019). https://doi.org/10.1109/TPAMI.2018.2798607
5. Benaissa, A., Retiat, B., Cebere, B., Belfedhal, A.E.: TenSEAL: a library for encrypted tensor operations using homomorphic encryption. CoRR abs/2104.03152 (2021). https://arxiv.org/abs/2104.03152

6. Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
7. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. ACM Trans. Comput. Theor. **6**(3), 13:1-13:36 (2014). https://doi.org/10.1145/2633600
8. Brutzkus, A., Gilad-Bachrach, R., Elisha, O.: Low latency privacy preserving inference. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 812–821. PMLR (2019). http://proceedings.mlr.press/v97/brutzkus19a.html
9. Cai, C., Sang, Y., Tian, H.: A multimodal differential privacy framework based on fusion representation learning. Connect. Sci. **34**(1), 2219–2239 (2022). https://doi.org/10.1080/09540091.2022.2111406
10. Cheon, J.H., Kim, A., Kim, M., Song, Y.: Homomorphic encryption for arithmetic of approximate numbers. In: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017. LNCS, vol. 10624, pp. 409–437. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70694-8_15
11. Chou, E.J., Gururajan, A., Laine, K., Goel, N.K., Bertiger, A., Stokes, J.W.: Privacy-preserving phishing web page classification via fully homomorphic encryption. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, 4–8 May 2020, pp. 2792–2796. IEEE (2020). https://doi.org/10.1109/ICASSP40776.2020.9053729
12. Deldjoo, Y., Schedl, M., Hidasi, B., Wei, Y., He, X.: Multimedia recommender systems: algorithms and challenges, 3rd edn. In: Recommender Systems Handbook (2020)
13. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
14. Fan, J., Vercauteren, F.: Somewhat practical fully homomorphic encryption. IACR Cryptol. ePrint Arch., 144 (2012). http://eprint.iacr.org/2012/144
15. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Mitzenmacher, M. (ed.) Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, 31 May–2 June 2009, pp. 169–178. ACM (2009). https://doi.org/10.1145/1536414.1536440
16. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K.E., Naehrig, M., Wernsing, J.: CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, JMLR Workshop and Conference Proceedings, ICML 2016, New York City, NY, USA, 19–24 June 2016, vol. 48, pp. 201–210. JMLR.org (2016). http://proceedings.mlr.press/v48/gilad-bachrach16.html
17. Halevi, S., Shoup, V.: Algorithms in HElib. IACR Cryptol. ePrint Arch. **2014**, 106 (2014)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90
19. Huynh, D.: Cryptotree: fast and accurate predictions on encrypted structured data. CoRR abs/2006.08299 (2020). https://arxiv.org/abs/2006.08299
20. McKeen, F., et al.: Innovative instructions and software model for isolated execution. In: Lee, R.B., Shi, W. (eds.) The Second Workshop on Hardware and Architectural Support for Security and Privacy, HASP 2013, Tel-Aviv, Israel, 23–24 June 2013, p. 10. ACM (2013). https://doi.org/10.1145/2487726.2488368
21. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, 28 June–2 July 2011, pp. 689–696. Omnipress (2011). https://icml.cc/2011/papers/399_icmlpaper.pdf
22. Rahulamathavan, Y.: Privacy-preserving similarity calculation of speaker features using fully homomorphic encryption. CoRR abs/2202.07994 (2022). https://arxiv.org/abs/2202.07994
23. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: a survey on recent advances and trends. IEEE Sig. Process. Mag. **34**(6), 96–108 (2017). https://doi.org/10.1109/MSP.2017.2738401
24. Rivest, R.L., Dertouzos, M.L.: On Data Banks and Privacy Homomorphisms (1978)
25. Microsoft SEAL (release 4.0). Microsoft Research, Redmond, WA, March 2022. https://github.com/Microsoft/SEAL
26. Sun, L., Wang, J., Zhang, K., Su, Y., Weng, F.: RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021, pp. 13860–13868. AAAI Press (2021). https://ojs.aaai.org/index.php/AAAI/article/view/17633
27. Wang, D., Xiong, D.: Efficient object-level visual context modeling for multimodal machine translation: masking irrelevant objects helps grounding. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021, pp. 2720–2728. AAAI Press (2021). https://ojs.aaai.org/index.php/AAAI/article/view/16376
28. Yao, A.C.: Protocols for secure computations (extended abstract). In: 23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3–5 November 1982, pp. 160–164. IEEE Computer Society (1982). https://doi.org/10.1109/SFCS.1982.38
29. Yu, F., et al.: ERNIE-ViL: knowledge enhanced vision-language representations through scene graphs. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021, pp. 3208–3216. AAAI Press (2021). https://ojs.aaai.org/index.php/AAAI/article/view/16431

30. Yu, W., et al.: CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020, pp. 3718–3727. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.343
31. Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021, pp. 10790–10797. AAAI Press (2021). https://ojs.aaai.org/index.php/AAAI/article/view/17289
32. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.: Tensor fusion network for multimodal sentiment analysis. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 9–11 September 2017, pp. 1103–1114. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/d17-1115
33. Zadeh, A., Zellers, R., Pincus, E., Morency, L.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016). https://doi.org/10.1109/MIS.2016.94