**硕士学位研究生 选题报告及文献总结**

论文题目 **Securing Private Transformer Inference Through Efficient MPC**

**指导教师：** 黄旗明副教授

**单 位** 计算机与通信工程学院通信工程系

**学 号：** M202161027

**作 者：** 马济 JIFAR

**专业名称：** 信息与通信工程

**入学时间：** 2022年09 月

2025年 9

**Abstract**

As the deployment of Transformer models for private inference becomes ubiquitous, concerns regarding the confidentiality of sensitive information processed by these models have intensified. Prior research has delved into secure inference methods for Transformer models through the utilization of secure multiparty computation (MPC), wherein the confidentiality of both model parameters and clients' prompts is upheld, existing frameworks are beset by constraints in terms of model performance, efficiency, and deployability. This research aimed to addresses this critical issue by proposing a methodology that leverages Secure Multi-Party Computation (MPC) to enhance the privacy of Transformer-based inference. The utilization of MPC allows multiple parties to jointly compute inference functions while maintaining the confidentiality of their respective inputs. The primary objective is to achieve robust privacy preservation without compromising computational efficiency. Efficiency and privacy are a central focus of this research, recognizing the inherent computational overhead associated with privacy-preserving techniques like MPC. By optimizing the MPC protocol for Transformer inference, we aim to strike a balance between privacy and computational speed, ensuring the practical viability of the proposed approach.

# Background

With the widespread adoption of Transformer models in natural language processing tasks, a critical concern has emerged surrounding the privacy of sensitive data processed during model inference. Traditional inference methodologies often entail a trade-off between model accuracy and the inadvertent exposure of confidential information, especially in applications dealing with personal, financial, or healthcare-related data. As the deployment of Transformer models continues to expand across diverse domains, addressing these privacy challenges becomes imperative to ensure responsible and ethical use of machine learning technologies.

Conventional inference processes, while effective in delivering accurate predictions, lack inherent mechanisms to safeguard against the potential compromise of private inputs. The inherently transparent nature of Transformer architectures exacerbates these privacy concerns, particularly in scenarios where the model is tasked with processing information of a sensitive nature. As a consequence, there is a growing need for innovative approaches that not only preserve the high-performance capabilities of Transformers but also ensure the privacy and security of sensitive data during the inference stage.

Multi-Party Computation (MPC) presents a promising avenue for reconciling the dual objectives of privacy preservation and efficient model inference. By enabling secure computation across multiple parties without revealing individual inputs, MPC offers a cryptographic solution to mitigate privacy risks associated with data exposure.

Previous research has delved into the realm of secure inference for Transformer models through the application of secure multiparty computation (MPC). In these studies, the focus will be on safeguarding the secrecy of both model parameters and clients' prompts or inputs. However, despite these efforts, the existing frameworks exhibit limitations concerning model performance, overall efficiency, and practical deployment. To improve the performance, working on high quality approximations for expensive functions, such as GeLU and Softmax, which significantly reduce the cost of secure inference while preserving the model performance will be the main target.

**Introduction**Top of Form

The ubiquity of Transformer models in natural language processing tasks has propelled advancements in various domains, from machine translation to sentiment analysis. However, the widespread use of these models raises critical concerns regarding the privacy of sensitive data processed during inference. Traditional inference methods often involve the exposure of confidential information, posing a significant risk in applications where data privacy is paramount.

In response to these challenges, this work introduces a pioneering approach to address the dual objectives of privacy and efficiency in Transformer model inference. Our focus is on leveraging Multi-Party Computation (MPC) to establish a secure environment for computations involving private data. MPC allows multiple parties to jointly compute a function over their inputs while keeping those inputs private. This cryptographic technique offers a promising solution to the privacy concerns associated with sensitive data inputs in the context of Transformer models.

The CRYPTGPU framework introduced a system that significantly enhances privacy-preserving machine learning by implementing all operations on the GPU. But these Crypt GPU only supports models with millions of parameters and not suitable for large models with billion parameters. Addressing the challenge of scaling privacy-preserving machine learning to larger models, MPCFORMER introduced an innovative solution leveraging Secure Multi-Party Computation (MPC) and Knowledge Distillation (KD)[2].Through rigorous evaluations, MPCFORMER demonstrated its ability to significantly accelerate Transformer inference in MPC settings while maintaining comparable performance to the original models. Evaluations conducted on the IMDb dataset showcased MPCFORMER achieving performance levels comparable to BERTBASE, with a remarkable 5.3× speedup. Furthermore, on the widely used GLUE benchmark, MPCFORMER achieved 97% of BERTBASE performance while boasting a 2.2× speedup. Surprisingly, PUMA frameworks [4] build upon the foundation laid by MPCFORMER, a state-of-the-art MPC framework recognized for its contributions to secure Transformer model inference, and showed great performance. Notably, the PUMA framework has demonstrated remarkable performance gains, boasting a speed that is approximately 2× faster than MPCFORMER. Furthermore, PUMA[ achieves this accelerated pace while maintaining a level of accuracy comparable to plaintext models, a feat that previous efforts in the field have struggled to accomplish. Of particular significance is PUMA's ability to evaluate the LLaMA-7B model in an impressively short timeframe, requiring only around 5 minutes to generate 1 token. This accomplishment represents a groundbreaking milestone, as it marks the first instance where a model of such substantial parameter size can be effectively evaluated under the constraints of MPC. While MPCFORMER introduced innovative techniques using MPC and Knowledge Distillation, PUMA extends this work by focusing on function approximations and secure procedures for Embedding and Layer Norm, and **thus providing a foundation for our approach.**

The goal of this research is to present a framework that ensures the confidentiality of private data during the inference process, without compromising computational efficiency. By adopting MPC, we aim to contribute to the growing body of privacy-preserving techniques for machine learning models, with a specific emphasis on the Transformer architecture.

**Transformer model**

The Transformer model is a type of neural network architecture introduced in the paper "Attention is All You Need" by Vaswani et al., published in 2017. This architecture has been widely adopted in natural language processing (NLP) and other machine learning tasks due to its effectiveness in handling sequential data. In particular, the two-stage training strategy for Transformer models has been shown to be effective in extensive settings and has become the domincated. In this training strategy, Transformer models are first pre-trained on a large dataset for general understanding, and then fine-tuned on a small downstream dataset to learn task-specific features. In this work, we consider this paradigm as the default setting, where we assume that model providers use pre-trained Transformer weights from elsewhere, and only have downstream data.

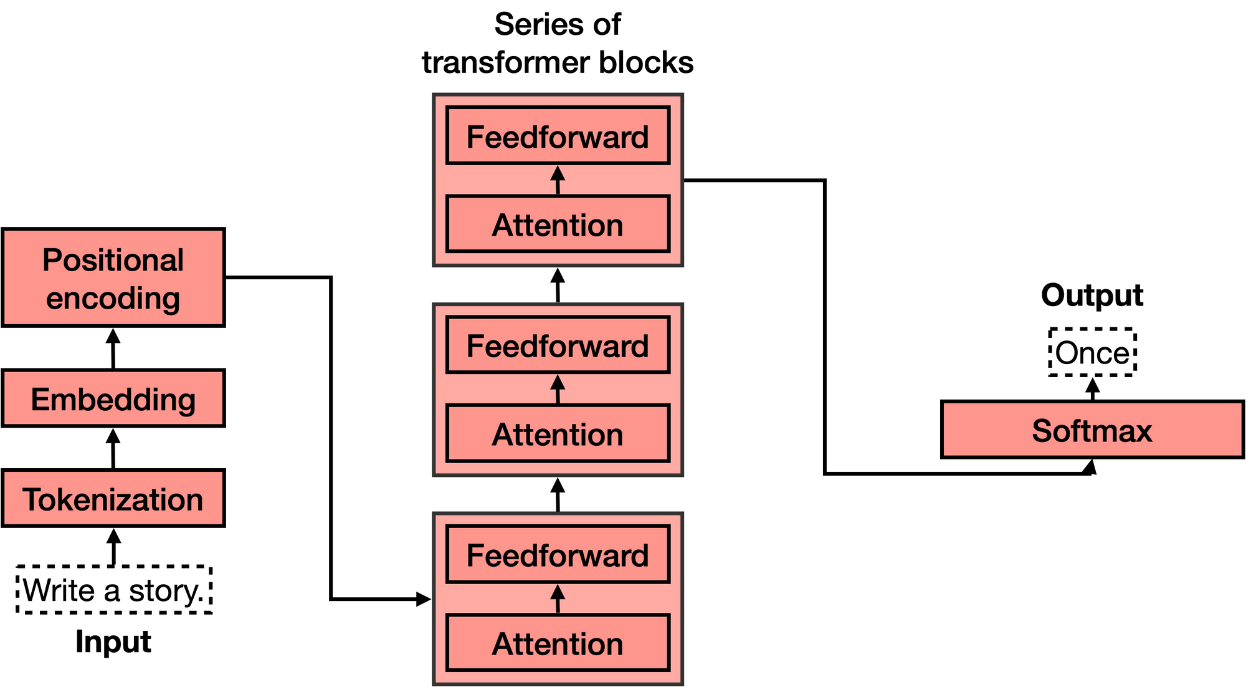


Figure 1 Transformer model Architecture

Key components of the Transformer model include:

1. **Self-Attention Mechanism:** This is a mechanism that allows the model to weigh the importance of different parts of the input sequence when making predictions at a particular position. Given inputs (Q, K, V), the Attention function is computed as Attention(Q, K, V) = Softmax(Q·KT + M{0,−∞}) · V, where M{0,−∞}, which is composed of {0, −∞}, is used to perform masking Attention in Decoder, and it can be viewed as a bias matrix. Besides, [Vaswani et al., 2017] proposed Multi-Head Attention to jointly attend to information from different representation subspaces at different positions. It enables the model to consider context from the entire input sequence, rather than relying solely on the previous or next elements in the sequence.
2. **Multi-Head Attention:** The self-attention mechanism is extended to multiple heads, each attending to different parts of the input. This allows the model to capture different aspects of the relationships within the sequence.
3. **Positional Encoding:** Since the Transformer model doesn't inherently understand the order of the input sequence, positional encodings are added to the input embeddings to provide information about the positions of tokens in the sequence.
4. **Encoder-Decoder Architecture:** The Transformer can be used for both encoder and decoder tasks. For sequence-to-sequence tasks (like machine translation), an encoder processes the input sequence, and a decoder generates the output sequence. Both the encoder and decoder consist of multiple layers of self-attention and feedforward neural networks.
5. **Feedforward Neural Network:** Each attention layer is followed by a feedforward neural network that processes the information captured by the attention mechanism. FFN is applied to each position separately and identically. This consists of two linear transformations with a activation in between, and the most common used activation function is GeLU. Given input x and parameters {W1, b1,W2, b2}, FFN can be formalized as FFN(x) = W2GeLU(W1x + b1) + b2,and the parameters of linear transformations are different from layer to layer.

The Transformer architecture has proven to be highly scalable and efficient, allowing for parallelization of training. It has become the foundation for many state-of-the-art models in NLP, such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and T5 (Text-to-Text Transfer Transformer), among others.

**Secure Multi-Party Computation**

Secure Multi-Party Computation (SMPC) is a cryptographic paradigm designed to facilitate collaborative computation among multiple parties while safeguarding the privacy of their individual inputs. The fundamental premise of SMPC lies in its ability to enable joint computations without exposing sensitive data to any participating entity. This cryptographic technique leverages various protocols, including homomorphic encryption, secret sharing, and secure function evaluation, to ensure that parties can collectively compute a function without divulging their raw input information.

As described by prior research, the concept of SMPC is rooted in the seminal paper "Secure Multi-Party Computation" by Andrew C. Yao, published in 1982. Yao's work laid the theoretical foundation for privacy-preserving computations by introducing the notion of secure function evaluation. Since then, advancements in cryptography have given rise to practical implementations of SMPC, addressing the increasing need for collaborative data analysis and privacy-preserving machine learning.

Homomorphic encryption, a key component of SMPC, allows computations to be conducted directly on encrypted data, ensuring that the original information remains confidential throughout the collaborative process. Additionally, secret sharing schemes play a crucial role by dividing a secret into shares distributed among the parties, ensuring that no single entity possesses complete knowledge of the original input. These cryptographic techniques collectively contribute to the overarching goal of SMPC: striking a balance between collaborative computation and individual data privacy.

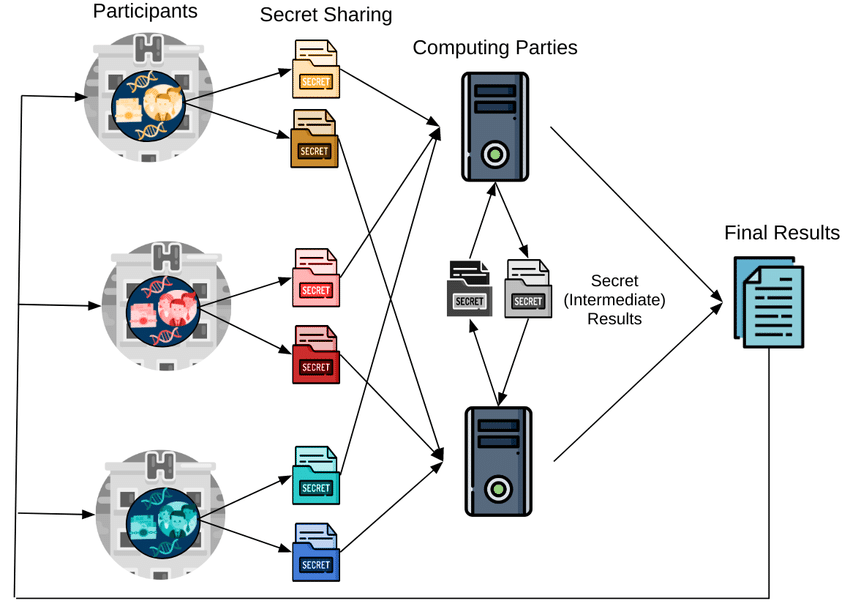


Figure 2 MPC

**Research significance**

SMPC finds application in diverse scenarios, including secure auctions, joint data analysis, and privacy-preserving machine learning models. While the technique offers robust privacy guarantees, challenges persist, primarily in terms of computational complexity. The efficiency of SMPC depends on the specific cryptographic protocols employed and the intricacy of the functions being computed.

As machine learning models, particularly the **potent Transformers**, become ubiquitous in diverse applications ranging from natural language processing to computer vision, the pressing need to safeguard the privacy of individuals whose data fuels these models has come to the forefront.

Transformers, renowned for their exceptional performance across various tasks, are frequently employed in inference scenarios where pre-trained models make predictions on new data. In this context, the security and privacy of the inference process become paramount, especially when handling data containing sensitive information.

The research's focus on leveraging Secure Multi-Party Computation (MPC) is a pivotal aspect of its significance. MPC, a cryptographic technique, facilitates collaborative computation of functions over private inputs from multiple parties. By applying MPC to secure transformer inference, the research aims to provide a robust cryptographic solution for preserving privacy during the computation of predictions.

However, the efficiency of such privacy-preserving techniques is a critical consideration. While MPC offers a powerful means of privacy protection, **it often introduces computational overhead.** Addressing this efficiency aspect is vital to ensure that the proposed methods do not compromise the speed and responsiveness of the inference process, making them feasible for practical deployment.

The real-world applications of this research are extensive, particularly in domains where privacy is paramount, such as healthcare, finance, and communication systems. The ability to protect sensitive information in these areas is not only essential for user trust but also aligns with regulatory requirements related to data protection.

Furthermore, the research will contributes to the broader field of privacy-preserving machine learning model specifically transformer. By proposing and validating methods tailored specifically for securing Transformer models in inference scenarios, it has the potential to set standards for privacy-preserving practices in machine learning. These advancements are poised to influence future research directions and impact industry applications, fostering a more secure and privacy-conscious integration of machine learning in various domains.

**Objectives**

1. **Enhancing Security in Transformer Inference:** The primary objective could be to develop and implement security measures to safeguard the inference process of Transformer models. This may involve addressing potential vulnerabilities, mitigating attacks, and ensuring the confidentiality and integrity of the model during the inference phase.
2. **Ensuring Privacy through MPC:** Leveraging MPC to enable secure computations across multiple parties without revealing sensitive information, thereby protecting the privacy of the data and the model being used for inference.
3. **Optimizing Efficiency in MPC:** Improving the efficiency of Multi-Party Computation for Transformer inference. This might involve developing techniques **specially deeply focus on exploring quantization method** to reduce computation overhead, communication costs, or latency associated with MPC, making it a more practical and scalable solution for securing private Transformer inference.

**Methodology**

**Literature Review:**

* Literature review is the most important part in carrying out quality research. To get better understanding of the concepts, leading published research papers related to MPC, Transformer, private inference will be deeply reviewed.

**MPC Integration:**

* Integrate MPC protocols into the Transformer inference pipeline to facilitate secure computation without revealing individual inputs.
* Select appropriate MPC techniques that align with the requirements of Transformer architectures.

**Efficiency Optimization:**

* Exploring optimization strategies, parallelization techniques, or hardware acceleration to maintain competitive performance.

# Expected Outcomes

**Privacy Preservation:**

* + Enhanced privacy for sensitive data used in Transformer model inference.
  + Mitigation of privacy risks associated with sharing or outsourcing computations to third parties.

**Secure Inference:**

* + Ensured confidentiality of model parameters and intermediate results during the inference process.
  + Protection against potential attacks aimed at extracting sensitive information from the model.

**Efficient MPC Implementation:**

* + Development of efficient MPC protocols tailored for Transformer models, minimizing computational overhead.
  + Optimization of communication and computation to enable practical and scalable secure inference.

**Performance Evaluation:**

* + Comparative analysis of the secured Transformer inference against baseline (non-secure) inference and some of the existing frameworks to assess the impact on performance.
  + Evaluation of trade-offs between privacy and computational efficiency.

**Regulatory Compliance:**

* Ensure compliance with data protection regulations and privacy laws, such as GDPR or HIPAA, depending on the jurisdiction and nature of the data. Adhering to these standards helps in safeguarding user privacy and avoiding legal complications.

**Challenges and limitations**

We will probably encounter several blocks while conducting our research. Due to its complexity and resource constraints, implementing large model will be challenging, and it require better understanding with high critical thinking. Some of the main challenges are listed below.

**Computational Overhead:**

* MPC involves complex cryptographic protocols and computations, leading to significant computational overhead. The efficiency of MPC protocols may be a concern, particularly when dealing with large Transformer models and extensive datasets.

**Communication Overhead:**

* The exchange of encrypted messages and secure computation across multiple parties introduces communication overhead. Transmitting data securely among parties can be resource-intensive.

**Model Size and Complexity:**

* Transformer models, especially those designed for state-of-the-art natural language processing tasks, can be large and complex. Securing the private inference of such models through MPC may be challenging due to the size of the model parameters.

**Resource Constraints:**

* MPC protocols often require significant computational resources. Implementing them in resource-constrained environments, such as edge devices, may be challenging.

**Conclusion**

In conclusion, the study on "Securing Private Transformer Inference Through Efficient MPC" underscores the critical intersection of privacy preservation and efficient computation in the realm of machine learning. Through the implementation of robust Multi-Party Computation (MPC) protocols tailored for Transformer models, the research aims to fortify the confidentiality of sensitive information during inference. The anticipated outcomes include enhanced privacy, secure inference processes, and the development of efficient protocols that strike a balance between privacy and computational efficiency.

**References**

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Li, Dacheng, et al. "MPCFormer: fast, performant and private Transformer inference with MPC." *arXiv preprint arXiv:2211.01452* (2022).

Ng, Lucien KL, and Sherman SM Chow. "SoK: Cryptographic Neural-Network Computation." *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023.

Dong, Y., Lu, W. J., Zheng, Y., Wu, H., Zhao, D., Tan, J., ... & Cheng, W. (2023). Puma: Secure inference of llama-7b in five minutes.