Privacy-preserving Paralinguistic Tasks

Francisco Teixeira, Alberto Abad, Isabel Trancoso

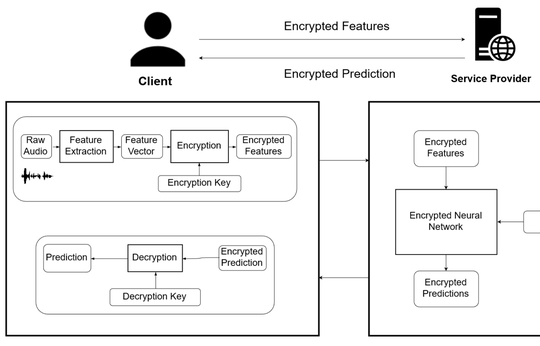
INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal

Abstract

Speech is one of the primary means of communication for humans. It can be viewed as a carrier for information on several levels as it conveys not only the meaning and intention predetermined by a speaker, but also paralinguistic and extralinguistic information about the speaker's age, gender, personality, emotional state, health state and affect. This makes it a particularly sensitive biometric, that should be protected. In this work we intent to explore how Leveled Homomorphic Encryption can be combined with a Neural Network to create a privacy-preserving machine learning framework for speech based health-related tasks. In particular, we will apply this framework to the detection and assessment of a Cold, Depression and Parkinson's Disease. Moreover, we will show how using a Quantized Neural Network, with discretized weights, allows us to apply a Leveled Homomorphic Encryption technique called *batching* that can be utilized to reduce the effective computational cost of this framework.

Motivation

- Speech contains a large amount of information about a person.
- Some of this information may be sensitive and not appropriate to be disclosed.
- Health related information taken from speech is especially sensitive, and should be protected.
- In speech it is common to have several samples for the same speaker.
- The use of *batching* is thus especially suited for speech-based applications.



Homomorphic Encryption

Concept

Mathematical operations can be computed on encrypted values (*ciphertexts*), yielding encrypted results:

$$Enc(a)+Enc(b)=Enc(a+b)$$
$$Enc(a)\times Enc(b)=Enc(a\times b)$$

Limitations

- Operations are limited to multiplications and additions (in most schemes).
- Performing homomorphic operations increases the amount of *noise* in a ciphertext, which can result in incorrect decryptions.
- The maximum amount of operations is determined by the encryption parameters.
- Changing the encryption parameters to perform more operations results in heavier computations.

Batching

- Levelled Homomorphic Encryption technique that allows several messages to be encrypted within the same ciphertext.
- Messages can be operated on as SIMD.
- Batching is incompatible with fractional encoding schemes.
- Encrypted values must be smaller than the plaintext modulus, to ensure correct decryptions.

Privacy-preserving Neural Networks (NNs)

- All operations in the NN are replaced by their HE counterparts [1].
- Non-linear Activation layers are replaced by polynomial approximations [2].
- Batch Normalization layer is introduced before each Activation, to ensure inputs fall within the convergence interval of the approximation [3].

Advantages

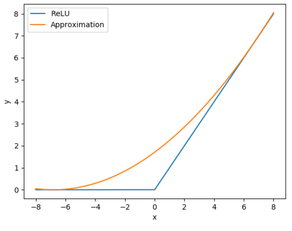
- Encrypted predictions can be computed over encrypted vectors of features.
- *Batching* can be used to perform several predictions at the same time.
- Provides a secure framework for both the user and the detainer of the model.

Disadvantages

- Predictions take much longer than in an unencrypted context (*s* vs *μs*).
- Network architecture is limited by noise growth, scaling and computational complexity.
- Polynomial approximations may result in less accurate models.
- Batching requires discretized weights and inputs, that may lead to accuracy degradation.

Proposed Solution

- Weight discretization through layer pre-computation and scaling.
- Input feature quantization.



Polynomial ReLU:

$$y = 0.037x^2 + 0.5x + 1.71$$

Neural Network Discretization:

$$\begin{cases} y_{FC} = A \cdot x + b \\ y_{BN} = \gamma \dfrac{(x-\mu)}{\sqrt{\sigma^2}} + \beta \\ y_{ACT} = ax^2 + bx + c \end{cases} \longrightarrow y_{FC+BN+ACT} = (A' \cdot x)^2 + b' \cdot x + c'$$

$$y_{FC+BN+ACT\ Scaled} = (\lfloor s * A' \rceil \cdot x)^2 + \lfloor s^2 * b' \rceil \cdot x + \lfloor s^2 * c' \rceil$$

μ-Law Quantization: $f(x) = sign(x)\dfrac{\ln(1+\mu|x|)}{\ln(1+\mu)}$ $\quad Q(x) = \left\lfloor \mu + \dfrac{f(x)+1}{2} + \dfrac{1}{2} \right\rfloor$

Experimental Setup

Datasets (Training and Development Corpus only)

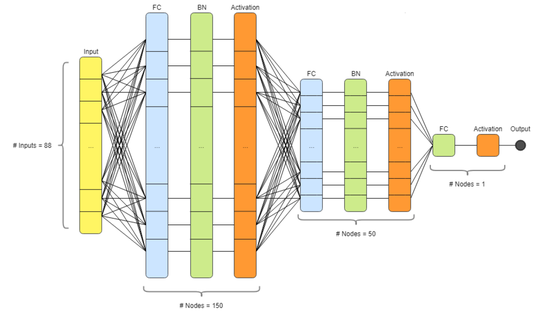
- Cold Corpus: URTIC – Classification.
- Depression Corpus: DAIC-WOZ – Classification and Regression.
- Parkinson's Disease Corpus : Spanish Corpus of UdeA – Regression.

Features:

- *Depression and Cold*: eGeMAPS features.
- *Parkinson's Disease:* Specialized feature set for PD, 36 GeMAPS based features plus 78 MFCC based features.

Encryption Parameters

- Polynomial Modulus: 16,384.
- Plaintext Modulus: Value larger than 2^{59}.
- Coefficient Modulus selected for a security level of 128 bits.



Results (4-bit Quantization, Scaling factor s = 150)

- 16,384 simultaneous predictions take around 23 seconds, using SEAL.
- The effective cost for a single prediction is around 1.4 milliseconds.

Method	F1 Score	Precision	Recall
NN	55.4	59.9	59.6
QNN	60.3	60.2	60.6
Scaled QNN	59.8	60.0	60.5

Method	RMSE	MAE
NN	6.69	5.59
QNN	6.74	5.62
Scaled QNN	6.67	5.63

Tables I,II – Results for Depression (Classification and Regression)

Method	F1 Score	Precision	Recall
NN	56.1	63.2	54.9
QNN	53.0	56.7	66.8
Scaled QNN	50.2	56.5	66.9

Table III – Results for Cold

Method	RMSE	MAE	ρ
NN	16.1	12.7	.43
QNN	15.9	12.6	.53
Scaled QNN	15.8	12.6	.52

Table IV – Results for PD

- **NN** – Baseline Neural Network with polynomial activation functions.
- **QNN** – Neural Network with polynomial activation function and quantized inputs.
- **Scaled QNN** – Equal to the QNN but with scaled weights.

Conclusions and Future Work

- This work showed that discretizing an NN and quantizing its inputs can be done with minimal accuracy degradation.
- Nonetheless, the proposed framework limits the size and architecture of NNs.
- As future work it would be interesting to explore other feature quantization functions.
- Additionally, *end-to-end* frameworks will also be explored, as these would remove the computational toll due to feature extraction from the client's side.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e Tecnologia (FCT) with reference UID/CEC/50021/2019.

References

[1] Gilad-Bachrach, R., Dowlin, N., Laine, K. et al., CryptoNets: "Applying Neural Networks to Encrypted Data with High Throughput and Accuracy" (2016).
[2] H. Chabanne, A. deWargny, J. Milgram, and C. Morel et al., "Privacy-Preserving Classification on Deep Neural Network." (2017).
[3] E. Hesamifard, H. Takabi, and M. Ghasemi, "CryptoDL: Deep Neural Networks over Encrypted Data." (2017).