universität innsbruck

Enclave-NN

Alexander Schlögl

Introduction

Neural Nets are universal approximators.

Technical Details

- Forward pass is a series of algebraic operations
- Fully defined by their architecture and weights (parameters)

We focus solely on inference phase, so NNs for us are static functions.

Problem Statement

Monetization requires keeping parameters private.

Problem Statement

Monetization requires keeping parameters private.

Current Method

Online Oracles

- parameters never public
- require sharing of data for inference
- require provider’s infrastructure

Problem Statement

Monetization requires keeping parameters private.

Current Method

Online Oracles

- parameters never public
- require sharing of data for inference
- require provider’s infrastructure

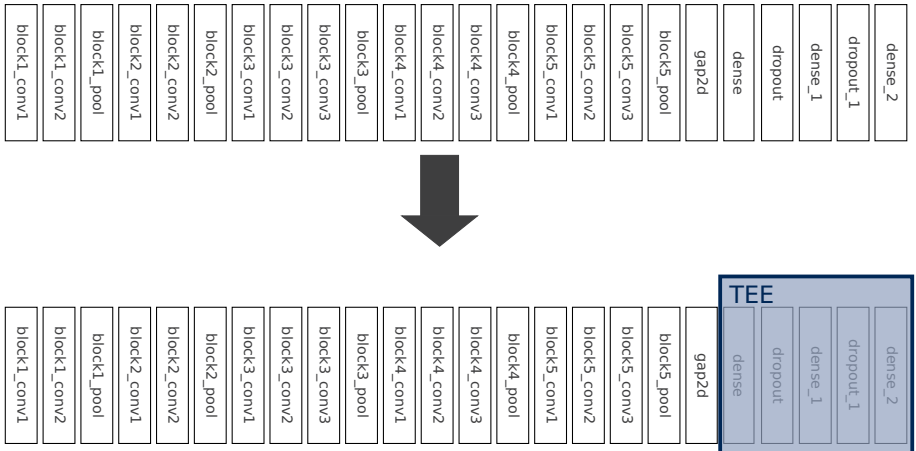
Idea

Use Trusted Execution Environments (TEEs) to hide parameters during inference

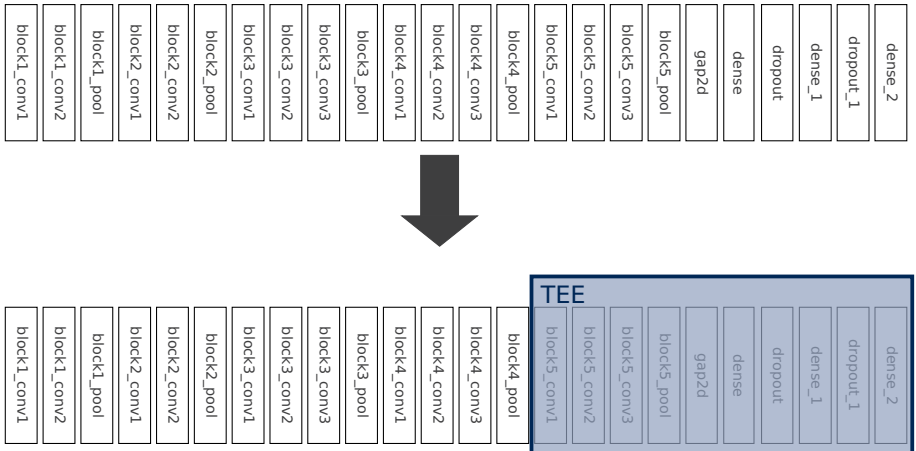
Approach


dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1

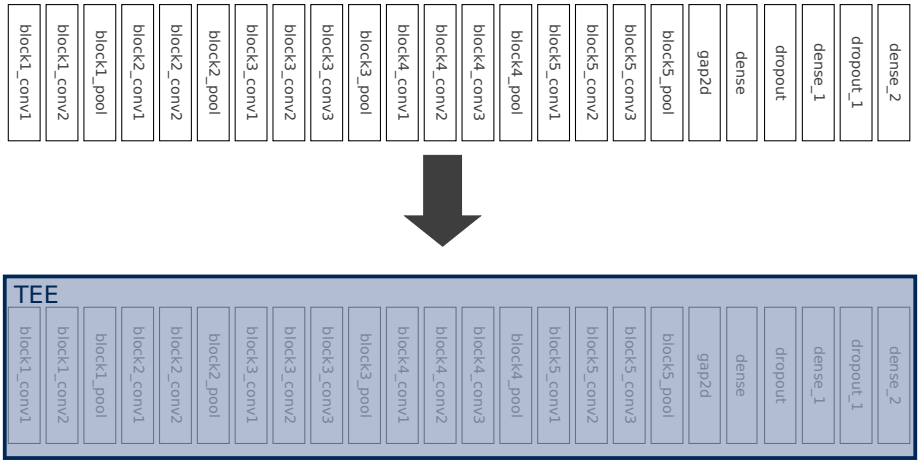
Approach


dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1
TEE
dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1

Approach


dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1
TEE
dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1

Approach


dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1
TEE
dense_2
dropout_1
dense_1
dropout
dense
gap2d
block5_pool
block5_conv3
block5_conv2
block5_conv1
block4_pool
block4_conv3
block4_conv2
block4_conv1
block3_pool
block3_conv3
block3_conv2
block3_conv1
block2_pool
block2_conv2
block2_conv1
block1_pool
block1_conv2
block1_conv1

Approach

Move last *n* layers into TEE, send protected model to user

Advantages
Less infrastructure required
Semi offline usage possible
Inference data can stay private

Disadvantages
Performance impact
Requires trust in manufacturer
Potentially larger attack surface

Evaluation Method

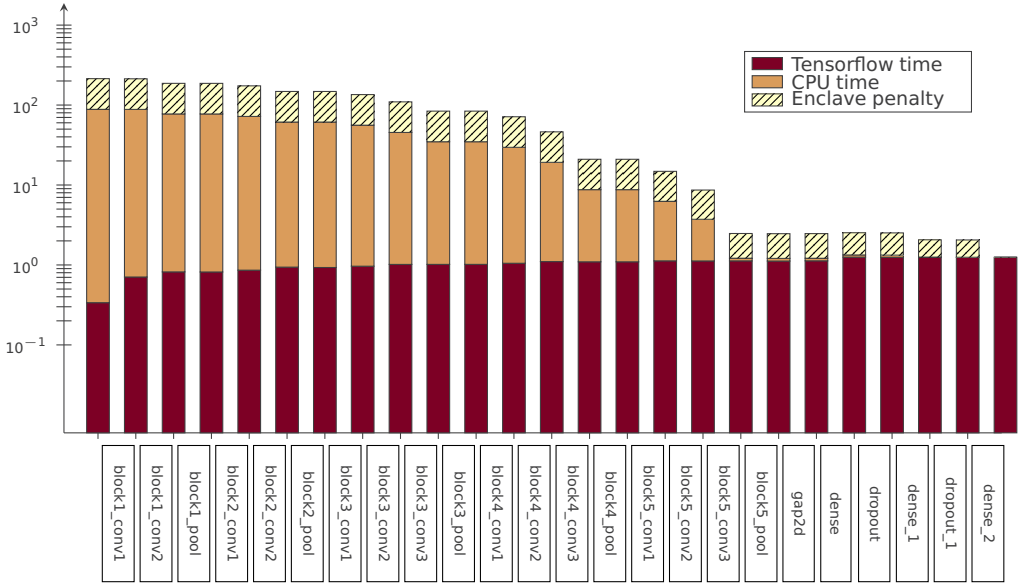
How large is the performance impact?

Procedure

1. Split NN
2. Compile TEE and native code
3. Measure inference time on single input
4. Separate CPU impact from TEE impact

Repeat for every possible split in NN

Results


10^3
10^2
10^1
10^0
10^{-1}
Tensorflow time
CPU time
Enclave penalty
block1_conv1
block1_conv2
block1_pool
block2_conv1
block2_conv2
block2_pool
block3_conv1
block3_conv2
block3_conv3
block3_pool
block4_conv1
block4_conv2
block4_conv3
block4_pool
block5_conv1
block5_conv2
block5_conv3
block5_pool
gap2d
dense
dropout
dense_1
dropout_1
dense_2

Open Problems

1. Architecture not hidden
2. No monetization prototype yet

MNIST Results

IMDB Results

Rotten Tomatoes Results