Stage 1: Data and Model Preparation

Output

Fullyconnected Layer

Flatten Layer

Convolution Layer (3 x3) Kernel size *valid* padding

MaxPooling Layer (3 x3) pool size

Input

28 x 28 x 1

(26, 26, 32)

(13, 13, 32)

5408

128

Correct Examples
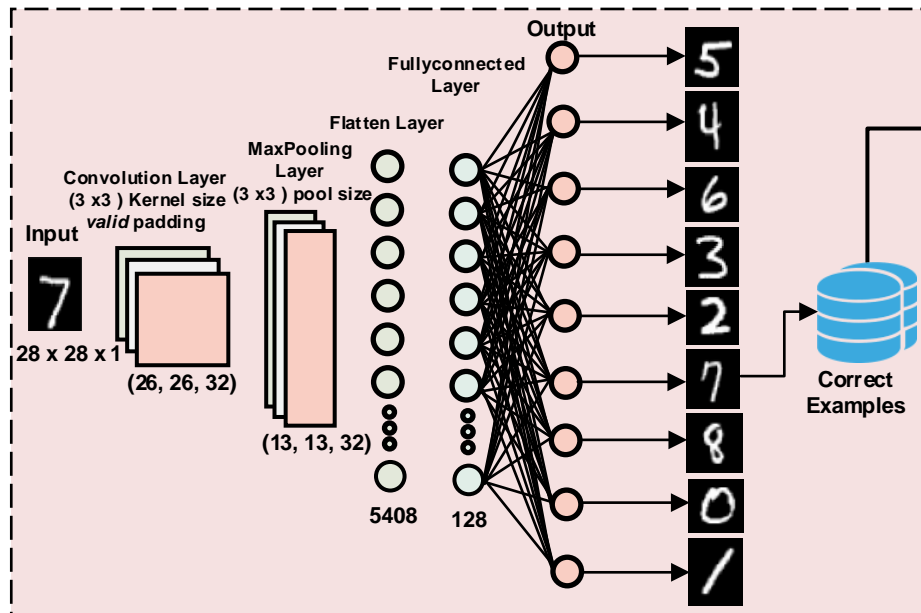
Stage 2: Adversarial Attack Simulation

Original Example

Attack Generation

Adversarial Example

PGD | BIM | FGSM | Deepfool

Original & Aversarial
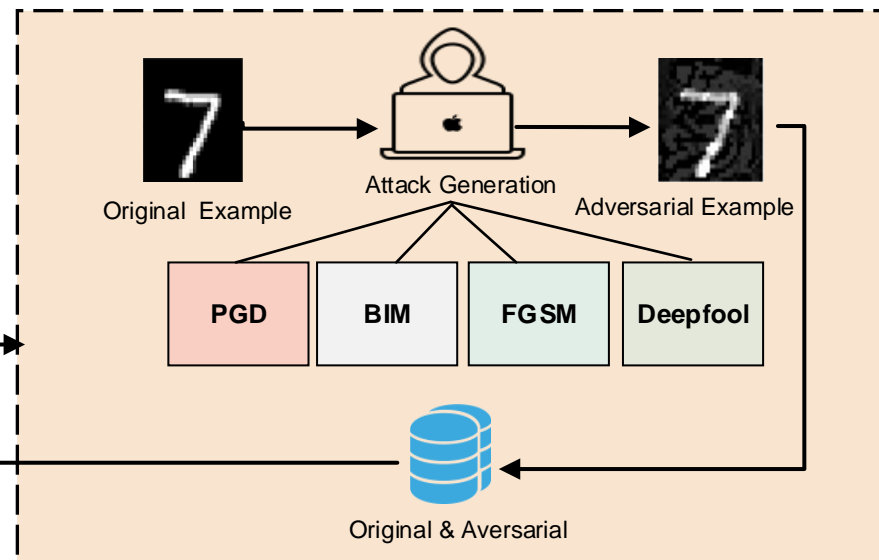
Stage 3: SHAP Analysis and Critical Pixel Visualization

Original/ Adversarial

Deep Explainer

SHAP

SHAP signatures

SHAP Signatures Data

Important Pixels (Attack Generation/Detection)
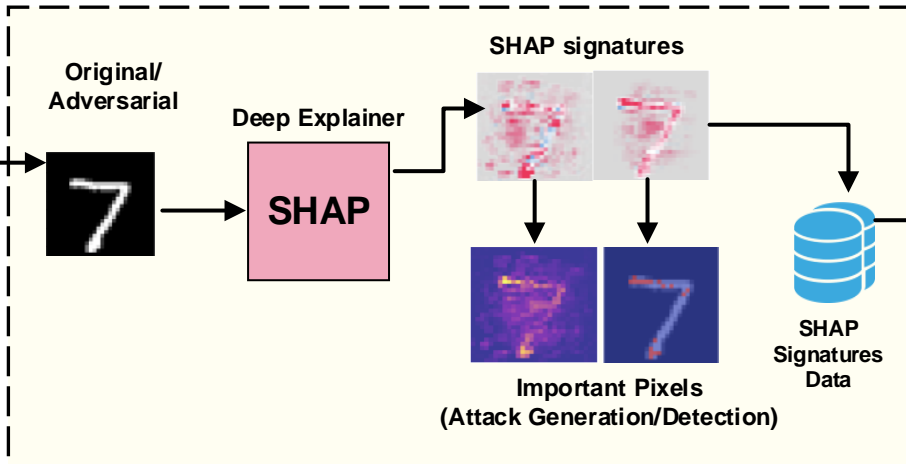
Stage 4: Binary Classification Using XAI Signatures

Flatten Layer

Fullycon nected Layer

Convolution Layer (3 x3) Kernel size *valid* padding

MaxPooling Layer (3 x3) pool size

Input

Output

Adversarial

1

Original

0

(26, 26, 32)

(13, 13, 32)

5408

128