Results

Substitution Targeted Attack

Target model: Default MNIST architecture from ZOO
Substitute model: 2 Dense layers
Simple model: 3 Convolutional layers, 2 Dense

Target model trained on training data
Substiute model trained on half of test set labeled by target model
Adversarial examples formed from other half of test set
All training = 5 epochs

Run 1

- Target model accuracy on clean data: 98.65
- Substitute model accuracy on clean data: 87.28
- SMNIST accuracy on clean data: 99.11
- Successes on sub: 22
- Successes on target: 22

Run 2

- Target model accuracy on clean data: 99.02
- Substitute model accuracy on clean data: 85.82
- SMNIST accuracy on clean data: 98.69
- Successes on sub: 24
- Successes on target: 24

Run 3

- Target model accuracy on clean data: 99.03
- Substitute model accuracy on clean data: 86.62
- SMNIST accuracy on clean data: 98.7
- Successes on sub: 25
- Successes on target: 25

Run 4

- Target model accuracy on clean data: 98.96
- Substitute model accuracy on clean data: 87.31
- SMNIST accuracy on clean data: 98.89
- Successes on sub: 25
- Successes on target: 25

Run 5

- Target model accuracy on clean data: 98.63

- Substitute model accuracy on clean data: 86.83
- SMNIST accuracy on clean data: 98.9
- Successes on sub: 23
- Successes on target: 23

Substitution Targeted Attack transfered to simpler model - 1

Target model: Default MNIST architecture from ZOO
Substitute model: 2 Dense layers
Simple model: 3 Convolutional layers, 2 Dense

Target model trained on training data
Substiute model trained on half of test set labeled by target model
Simple model trained on training data
Adversarial examples formed from other half of test set
All training = 5 epochs
Training shuffle off for consistency

Run 1

- Target model accuracy on clean data: 98.44
- Substitute model accuracy on clean data: 87.86
- SMNIST accuracy on clean data: 98.7
- Successes on sub: 25
- Successes on target: 25
- Successes on simple: 18
    - Correct target on simple: 18

Run 2

- Target model accuracy on clean data: 98.31
- Substitute model accuracy on clean data: 87.32
- SMNIST accuracy on clean data: 98.63
- Successes on sub: 24
- Successes on target: 24
- Successes on simple: 17
    - Correct target on simple: 16

Run 3

- Target model accuracy on clean data: 98.53
- Substitute model accuracy on clean data: 87.33
- SMNIST accuracy on clean data: 98.62
- Successes on sub: 27
- Successes on target: 27
- Successes on simple: 18
    - Correct target on simple: 18

Run 4

- Target model accuracy on clean data: 98.63
- Substitute model accuracy on clean data: 87.05
- SMNIST accuracy on clean data: 98.63
- Successes on sub: 27
- Successes on target: 27
- Successes on simple: 14
    - Correct target on simple: 14

Run 5

- Target model accuracy on clean data: 98.4
- Substitute model accuracy on clean data: 86.72
- SMNIST accuracy on clean data: 98.8
- Successes on sub: 23
- Successes on target: 23
- Successes on simple: 20
    - Correct target on simple: 19

Run 6

- Target model accuracy on clean data: 98.22
- Substitute model accuracy on clean data: 87.17
- SMNIST accuracy on clean data: 98.67
- Successes on sub: 24
- Successes on target: 24
- Successes on simple: 14
    - Correct target on simple: 0

Run 7

- Target model accuracy on clean data: 98.61
- Substitute model accuracy on clean data: 87.26
- SMNIST accuracy on clean data: 98.58
- Successes on sub: 20
- Successes on target: 20
- Successes on simple: 17
    - Correct target on simple: 0

Run 8

- Target model accuracy on clean data: 98.71
- Substitute model accuracy on clean data: 83.33
- SMNIST accuracy on clean data: 98.55
- Successes on sub: 15
- Successes on target: 15
- Successes on simple: 10
    - Correct target on simple: 0

Run 9

- Target model accuracy on clean data: 98.54
- Substitute model accuracy on clean data: 87.64
- SMNIST accuracy on clean data: 98.49
- Successes on sub: 29
- Successes on target: 29
- Successes on simple: 19
    - Correct target on simple: 19

Run 10

- Target model accuracy on clean data: 98.57
- Substitute model accuracy on clean data: 86.96
- SMNIST accuracy on clean data: 98.61
- Successes on sub: 27
- Successes on target: 27
- Successes on simple: 15
    - Correct target on simple: 0

Substitution Targeted Attack transfered to simpler model - 2

Target model: Default MNIST architecture from ZOO
Substitute model: 2 Dense layers
Simple model: 2 Convolutional layers, 2 Dense

Target model trained on training data
Substiute model trained on half of test set labeled by target model
Simple model trained on training data
Adversarial examples formed from other half of test set
All training = 5 epochs
Training shuffle off for consistency

Run 1

- Target model accuracy on clean data: 98.68
- Substitute model accuracy on clean data: 86.26
- SMNIST accuracy on clean data: 98.48
- Successes on sub: 20
- Successes on target: 20
- Successes on simple: 17
    - Correct target on simple: 17

Run 2

- Target model accuracy on clean data: 98.47
- Substitute model accuracy on clean data: 86.29
- SMNIST accuracy on clean data: 98.26
- Successes on sub: 23
- Successes on target: 23

- Successes on simple: 16
    - Correct target on simple: 0

Run 3

- Target model accuracy on clean data: 98.69
- Substitute model accuracy on clean data: 85.8
- SMNIST accuracy on clean data: 98.08
- Successes on sub: 22
- Successes on target: 22
- Successes on simple: 13
    - Correct target on simple: 0

Run 4

- Target model accuracy on clean data: 98.61
- Substitute model accuracy on clean data: 85.89
- SMNIST accuracy on clean data: 98.48
- Successes on sub: 29
- Successes on target: 29
- Successes on simple: 19
    - Correct target on simple: 1

Run 5

- Target model accuracy on clean data: 98.9
- Substitute model accuracy on clean data: 85.8
- SMNIST accuracy on clean data: 98.42
- Successes on sub: 28
- Successes on target: 28
- Successes on simple: 15
    - Correct target on simple: 0

Run 6

- Target model accuracy on clean data: 98.74
- Substitute model accuracy on clean data: 87.12
- SMNIST accuracy on clean data: 98.39
- Successes on sub: 21
- Successes on target: 21
- Successes on simple: 18
    - Correct target on simple: 17

Run 7

- Target model accuracy on clean data: 98.45
- Substitute model accuracy on clean data: 86.29
- SMNIST accuracy on clean data: 98.14
- Successes on sub: 27

- Successes on target: 27
- Successes on simple: 18
    - Correct target on simple: 0

Run 8

- Target model accuracy on clean data: 98.45
- Substitute model accuracy on clean data: 87.14
- SMNIST accuracy on clean data: 98.29
- Successes on sub: 29
- Successes on target: 29
- Successes on simple: 18
    - Correct target on simple: 0

Run 9

- Target model accuracy on clean data: 98.59
- Substitute model accuracy on clean data: 88.08
- SMNIST accuracy on clean data: 98.54
- Successes on sub: 21
- Successes on target: 21
- Successes on simple: 17
    - Correct target on simple: 17

Run 10

- Target model accuracy on clean data: 98.52
- Substitute model accuracy on clean data: 86.98
- SMNIST accuracy on clean data: 97.84
- Successes on sub: 28
- Successes on target: 28
- Successes on simple: 18
    - Correct target on simple: 0

Substitution Targeted Attack transfered to simpler model - 3

Target model: Default MNIST architecture from ZOO
Substitute model: 2 Dense layers
Simple model: 1 Convolutional layers, 2 Dense

Target model trained on training data
Substiute model trained on half of test set labeled by target model
Simple model trained on training data
Adversarial examples formed from other half of test set
All training = 5 epochs
Training shuffle off for consistency

Run 1

- Target model accuracy on clean data: 98.53
- Substitute model accuracy on clean data: 86.55
- SMNIST accuracy on clean data: 98.15
- Successes on sub: 28
- Successes on target: 28
- Successes on simple: 20
    - Correct target on simple: 0

Run 2

- Target model accuracy on clean data: 98.39
- Substitute model accuracy on clean data: 87.41
- SMNIST accuracy on clean data: 98.21
- Successes on sub: 28
- Successes on target: 28
- Successes on simple: 16
    - Correct target on simple: 16

Run 3

- Target model accuracy on clean data: 98.35
- Substitute model accuracy on clean data: 87.01
- SMNIST accuracy on clean data: 98.23
- Successes on sub: 26
- Successes on target: 26
- Successes on simple: 20
    - Correct target on simple: 0

Run 4

- Target model accuracy on clean data: 98.52
- Substitute model accuracy on clean data: 87.11
- SMNIST accuracy on clean data: 97.79
- Successes on sub: 23
- Successes on target: 23
- Successes on simple: 18
    - Correct target on simple: 0

Run 5

- Target model accuracy on clean data: 85.52
- Substitute model accuracy on clean data: 87.5
- SMNIST accuracy on clean data: 98.11
- Successes on sub: 23
- Successes on target: 23
- Successes on simple: 18
    - Correct target on simple: 0

Run 6

- Target model accuracy on clean data: 98.47
- Substitute model accuracy on clean data: 86.04
- SMNIST accuracy on clean data: 98.05
- Successes on sub: 30
- Successes on target: 30
- Successes on simple: 16
    - Correct target on simple: 0

Run 7

- Target model accuracy on clean data: 98.49
- Substitute model accuracy on clean data: 86.46
- SMNIST accuracy on clean data: 98.16
- Successes on sub: 23
- Successes on target: 23
- Successes on simple: 19
    - Correct target on simple: 19

Run 8

- Target model accuracy on clean data: 98.48
- Substitute model accuracy on clean data: 87.63
- SMNIST accuracy on clean data: 98.08
- Successes on sub: 26
- Successes on target: 26
- Successes on simple: 15
    - Correct target on simple: 15

Run 9

- Target model accuracy on clean data: 98.55
- Substitute model accuracy on clean data: 87.64
- SMNIST accuracy on clean data: 98.23
- Successes on sub: 29
- Successes on target: 29
- Successes on simple: 18 (1 initially wrong)
    - Correct target on simple: 18

Run 10

- Target model accuracy on clean data: 98.44
- Substitute model accuracy on clean data: 86.14
- SMNIST accuracy on clean data: 98.22
- Successes on sub: 23
- Successes on target: 23
- Successes on simple: 16
    - Correct target on simple: 0

Substitution Targeted Attack transfered to simpler model - 4

Target model: Default MNIST architecture from ZOO
Substitute model: 2 Dense layers
Simple model: 0 Convolutional layers, 2 Dense

Target model trained on training data
Substiute model trained on half of test set labeled by target model
Simple model trained on training data
Adversarial examples formed from other half of test set
All training = 5 epochs
Training shuffle off for consistency

Run 1

- Target model accuracy on clean data: 98.37
- Substitute model accuracy on clean data: 86.84
- SMNIST accuracy on clean data: 95.99
- Successes on sub: 22
- Successes on target: 22
- Successes on simple: 17
    - Correct target on simple: 0

Run 2

- Target model accuracy on clean data: 98.47
- Substitute model accuracy on clean data: 87.
- SMNIST accuracy on clean data: 96.21
- Successes on sub: 24
- Successes on target: 24
- Successes on simple: 17
    - Correct target on simple: 17

Run 3

- Target model accuracy on clean data: 98.15
- Substitute model accuracy on clean data: 85.89
- SMNIST accuracy on clean data: 96.1
- Successes on sub: 24
- Successes on target: 24
- Successes on simple: 17
    - Correct target on simple: 0

Run 4

- Target model accuracy on clean data: 98.64
- Substitute model accuracy on clean data: 86.1
- SMNIST accuracy on clean data: 96.06
- Successes on sub: 24
- Successes on target: 24
- Successes on simple: 19

    - Correct target on simple: 19

Run 5

- Target model accuracy on clean data: 98.54
- Substitute model accuracy on clean data: 86.05
- SMNIST accuracy on clean data: 96.05
- Successes on sub: 26
- Successes on target: 26
- Successes on simple: 19
    - Correct target on simple: 19

Run 6

- Target model accuracy on clean data: 98.71
- Substitute model accuracy on clean data: 85.93
- SMNIST accuracy on clean data: 95.92
- Successes on sub: 29
- Successes on target: 29
- Successes on simple: 17
    - Correct target on simple: 17

Run 7

- Target model accuracy on clean data: 98.53
- Substitute model accuracy on clean data: 86.43
- SMNIST accuracy on clean data: 96.12
- Successes on sub: 24
- Successes on target: 24
- Successes on simple: 17
    - Correct target on simple: 17

Run 8

- Target model accuracy on clean data: 98.61
- Substitute model accuracy on clean data: 85.21
- SMNIST accuracy on clean data: 96.15
- Successes on sub: 32
- Successes on target: 32
- Successes on simple: 17
    - Correct target on simple: 17

Run 9

- Target model accuracy on clean data: 98.38
- Substitute model accuracy on clean data: 86.47
- SMNIST accuracy on clean data: 95.9
- Successes on sub: 31
- Successes on target: 31

- Successes on simple: 17
    - Correct target on simple: 17

Run 10

- Target model accuracy on clean data: 98.43
- Substitute model accuracy on clean data: 86.66
- SMNIST accuracy on clean data: 96.08
- Successes on sub: 20
- Successes on target: 20
- Successes on simple: 17
    - Correct target on simple: 17