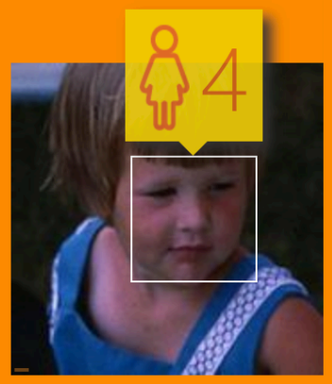
Attacks against Neural Networks

Student: Martin Matak
Supervisor: Georg Weissenbacher

How-Old.net

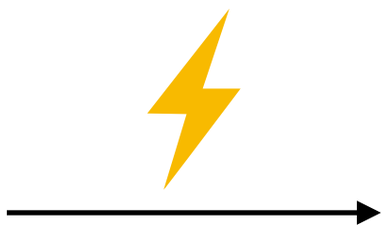
How old do I look? #HowOldRobot



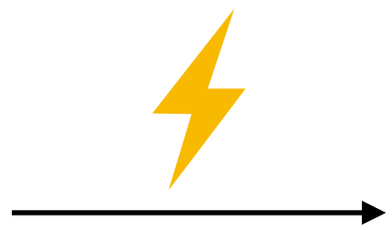
Sorry if we didn't quite get it right - we are still improving this feature..

Try Another Photo!

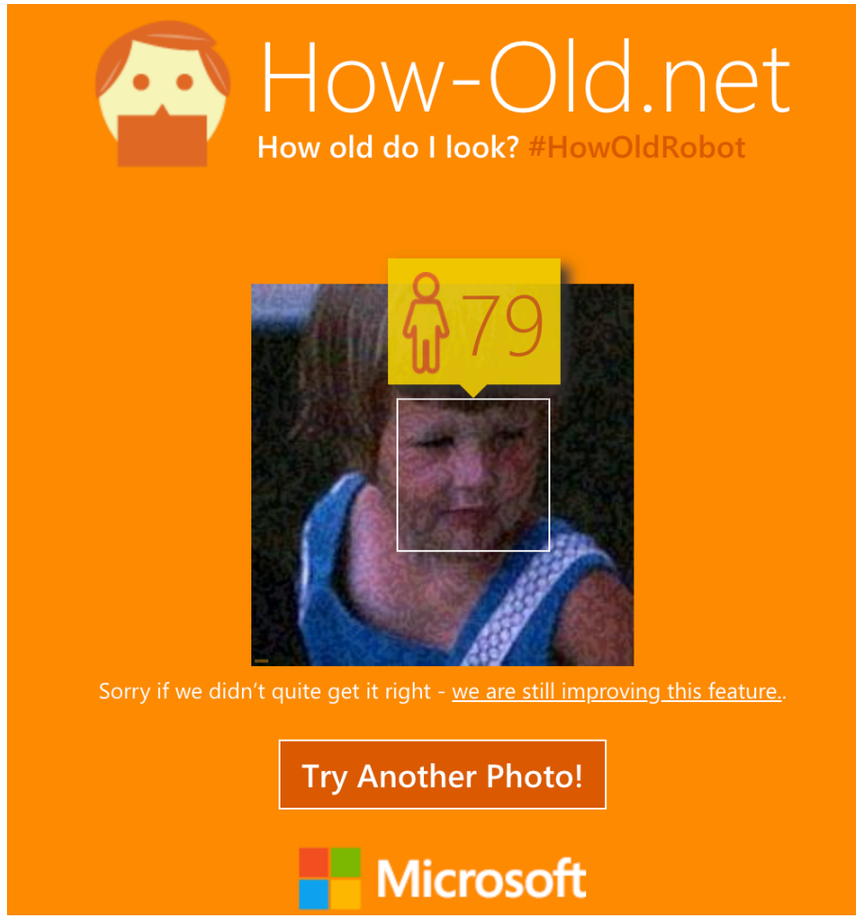
Microsoft




How-Old.net
How old do I look? #HowOldRobot
4
Sorry if we didn't quite get it right - we are still improving this feature..
Try Another Photo!
Microsoft




How-Old.net
How old do I look? #HowOldRobot
4
Sorry if we didn't quite get it right - we are still improving this feature..
Try Another Photo!
Microsoft


How-Old.net
How old do I look? #HowOldRobot
79
Sorry if we didn't quite get it right - we are still improving this feature..
Try Another Photo!
Microsoft

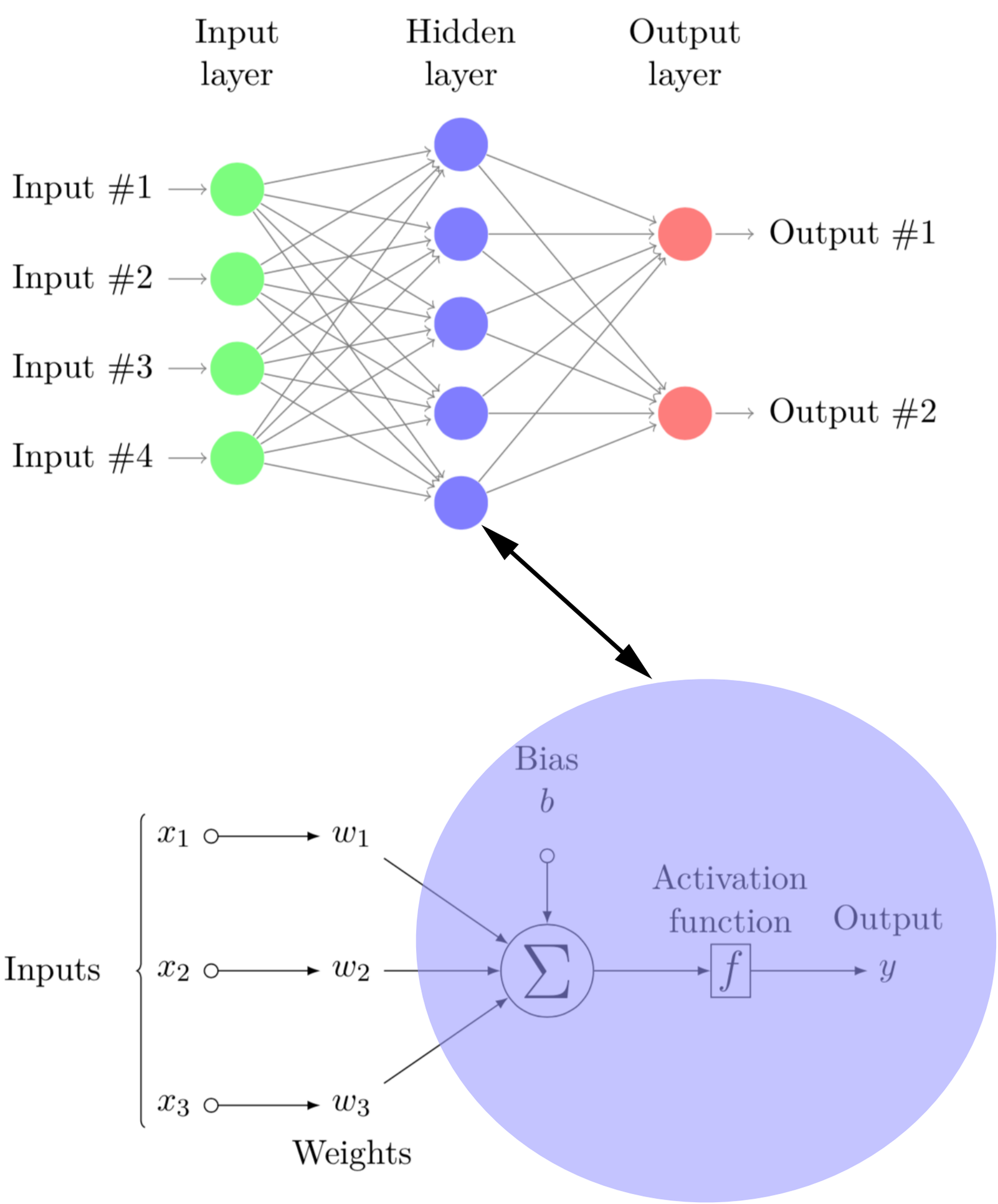
Outline:

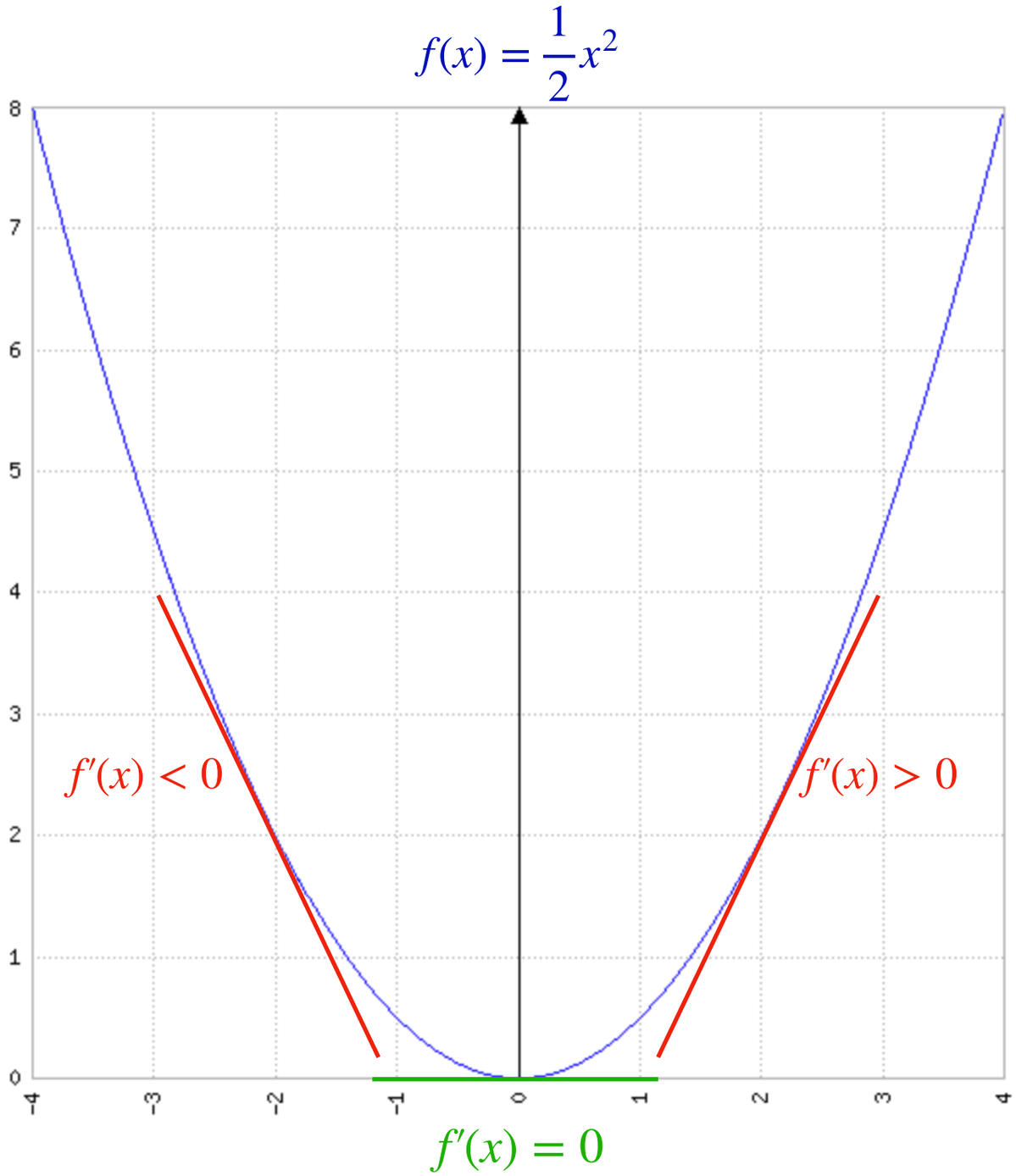
1. **Train neural networks for age estimation**

2. **Attack the networks**
   - **White-box environment**
   - **Black-box environment**

3. **The Semi-targeted Approach**

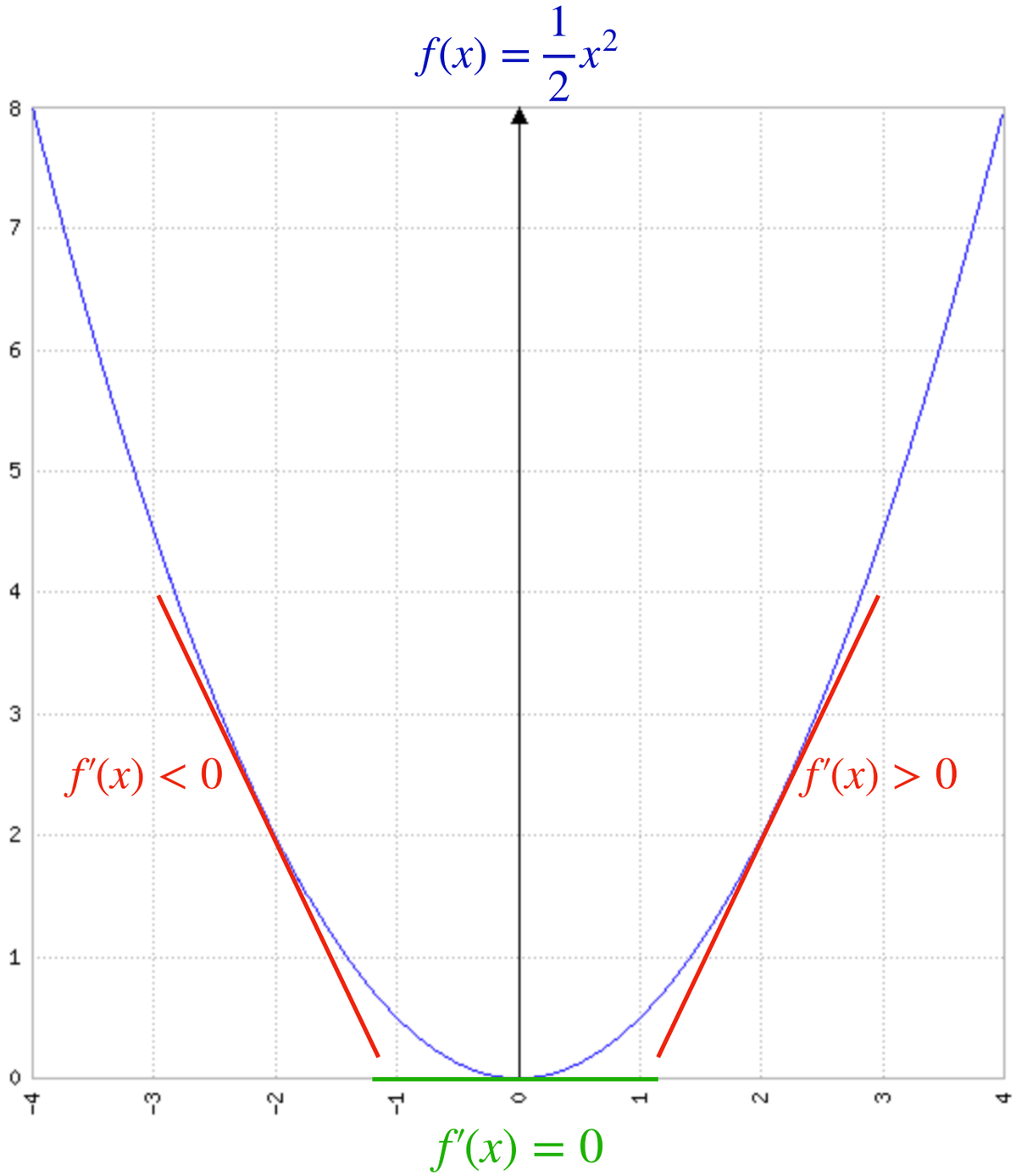
Background


Input layer
Hidden layer
Output layer
Input #1
Input #2
Input #3
Input #4
Output #1
Output #2
Bias
b
Activation function
Output
y
f
$\sum$
Inputs
x_1
x_2
x_3
w_1
w_2
w_3
Weights



Gradient descent algorithm

```
input: α,p

while true:

  gradient := f'(p)

  if gradient is small enough:
    break
  else:
    p := p - α * gradient
```

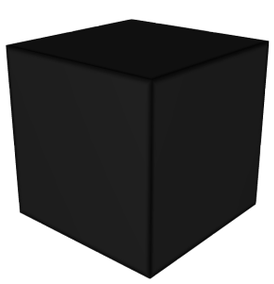


Gradient descent algorithm

```
input: α,p

while true:

  gradient := f'(p)

  if gradient is small enough:
    break
  else:
    p := p - α * gradient
```

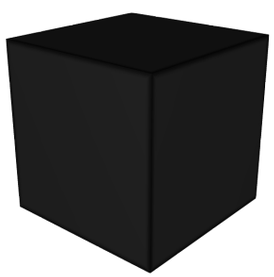
$$L(weights) = \frac{1}{S}\sum_{s=1}^{S}(y_s - neural\_network(x_s; weights))^2$$

Current classification



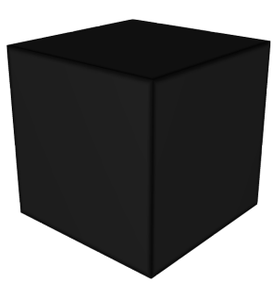
4

Misclassification



~~4~~

Targeted misclassification



90

Trained models

Dataset: 30,843 images

	Architecture	Optimizer	Validation MAE
model 1	ResNet-50	SGD	5.15
model 2	ResNet-50	Adam	6.77
model 3	InceptionResNetV2	SGD	4.50
model 4	InceptionResNetV2	Adam	3.92

White-box Attacks

Fast gradient sign method (FGSM)

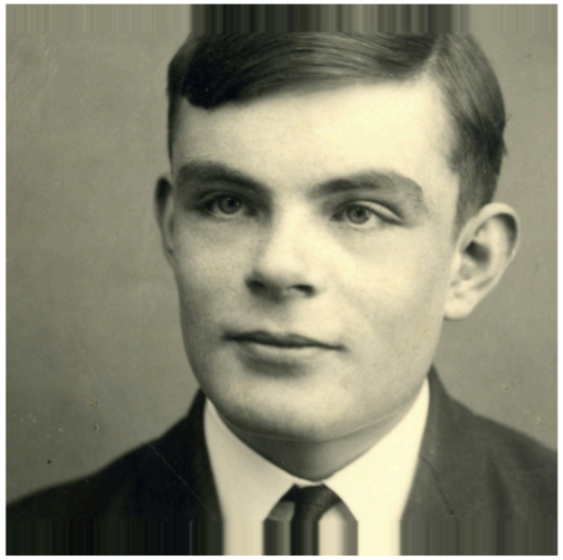
Fast gradient sign method (FGSM)

$$L(weights, x, target\_y) = (target\_y - neural\_network(x; weights))^2$$

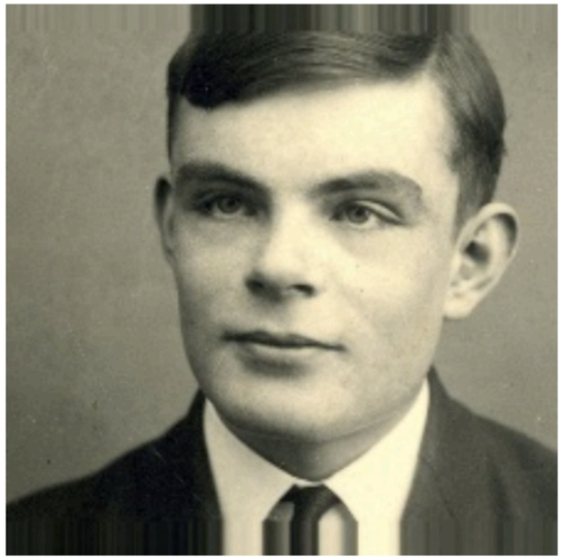
Fast gradient sign method (FGSM)

$$L(weights, x, target\_y) = (target\_y - neural\_network(x; weights))^2$$

```
input: x, target_y, α, neural network (white-box)

gradient := L'(weights, x, target_y)
       x := x - α * sign(gradient)
```



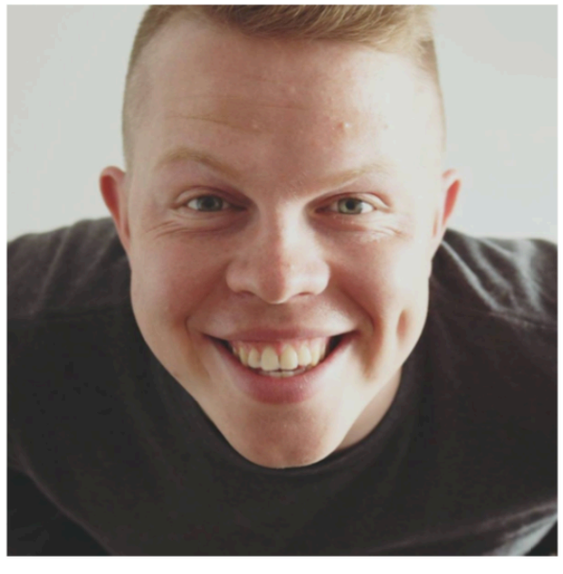
28



59

59

28



28

64

Carlini and Wagner (CW)

Carlini and Wagner (CW)

1. *Minimize* added perturbation
2. Must be adversarial (target class)
3. Must be a valid image

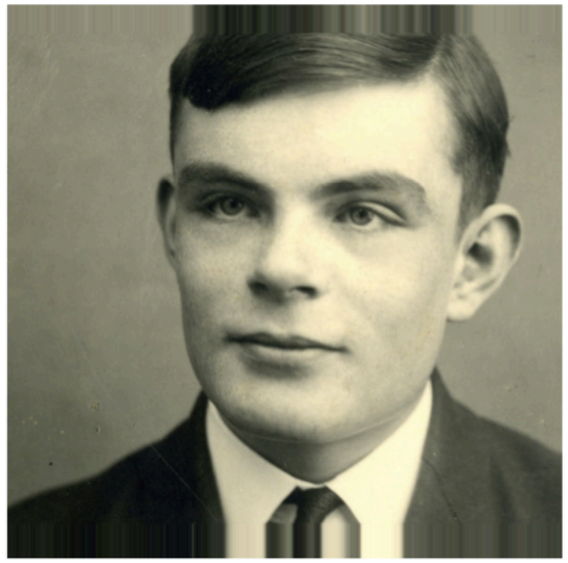
Carlini and Wagner (CW)

1. *Minimize* added perturbation

2. Must be adversarial (target class)

3. Must be a valid image

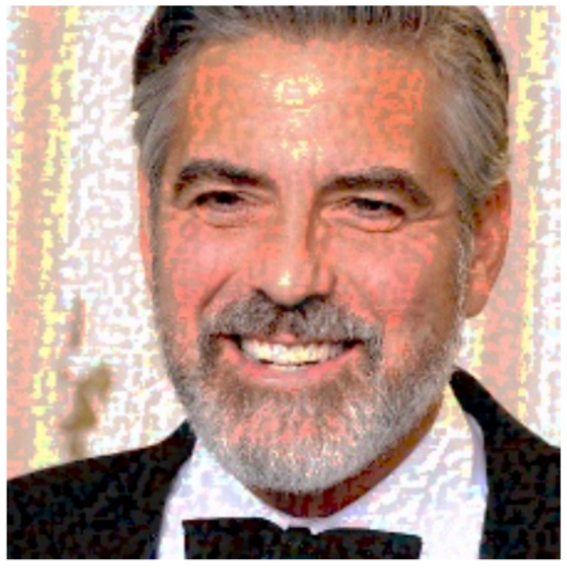
$$\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta})$$
$$\text{such that } \mathcal{C}(\boldsymbol{x} + \boldsymbol{\delta}) = t$$
$$\boldsymbol{x} + \boldsymbol{\delta} \in [0, 1]^n$$



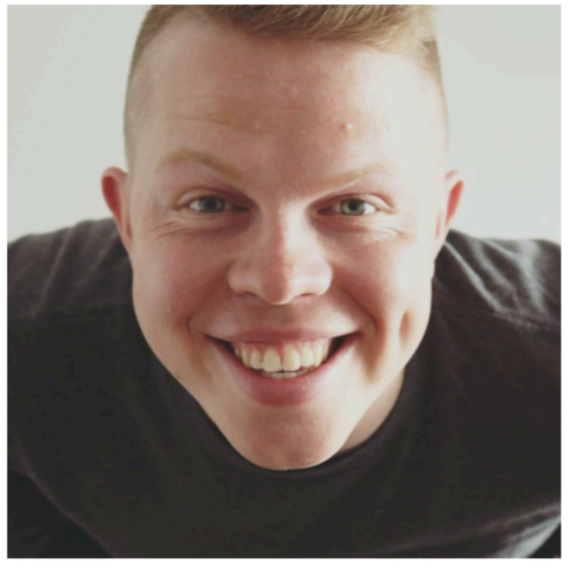
36

79

55



10



35

84

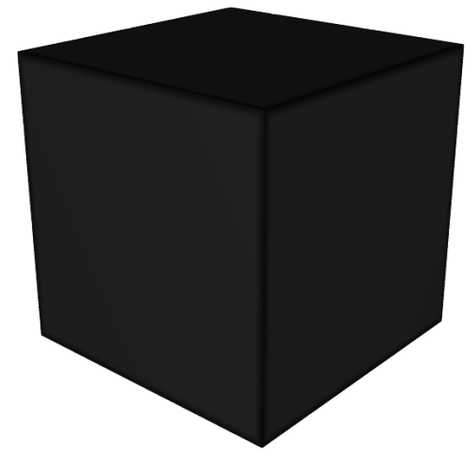
White-box: results

	clean samples MAE	FGSM MAE	FGSM AVG L2	CW MAE	CW AVG L2
model 1	6.69	32.25	1879.63	29.35	2977.28
model 2	9.50	**47.79**	1879.49	9.50	**0.00**
model 3	5.40	19.32	2510.92	7.51	1307.51
model 4	4.62	15.03	2511.10	6.79	1104.78

Black-box attacks

Adversary

Black-box Neural Network

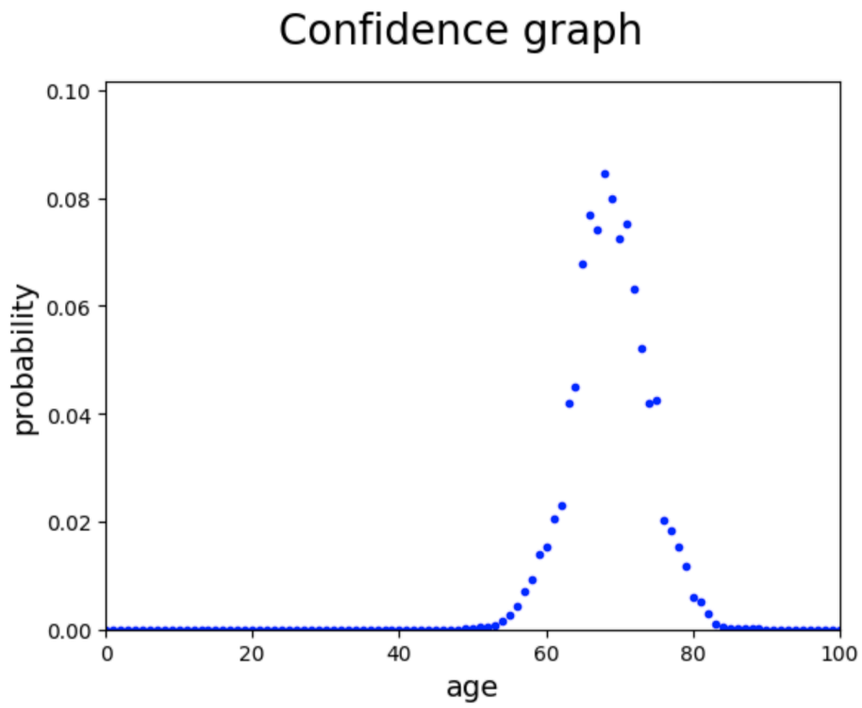


42

Boundary attack

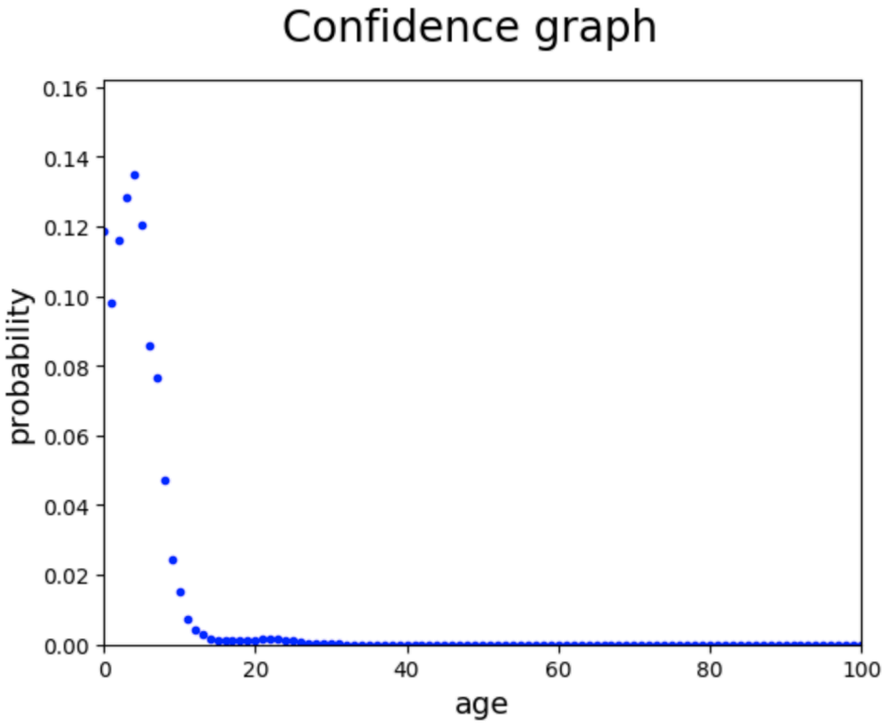
The starting image:

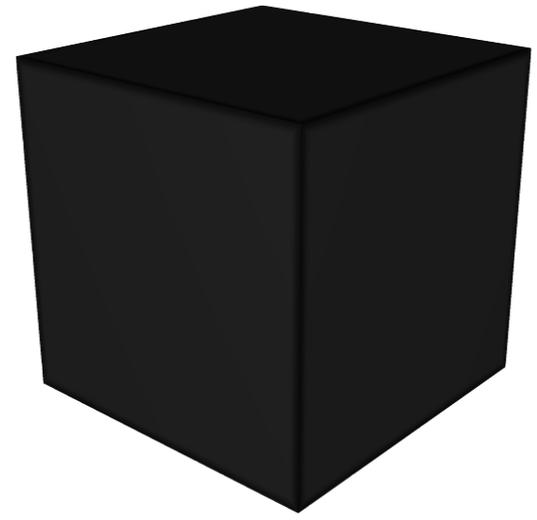
68

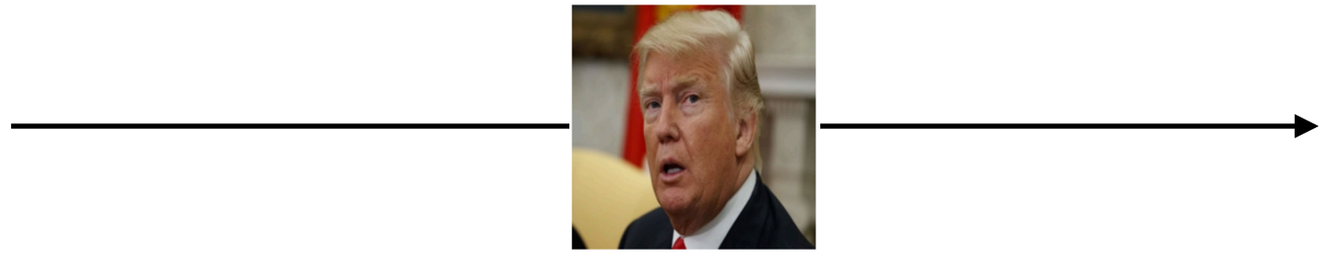


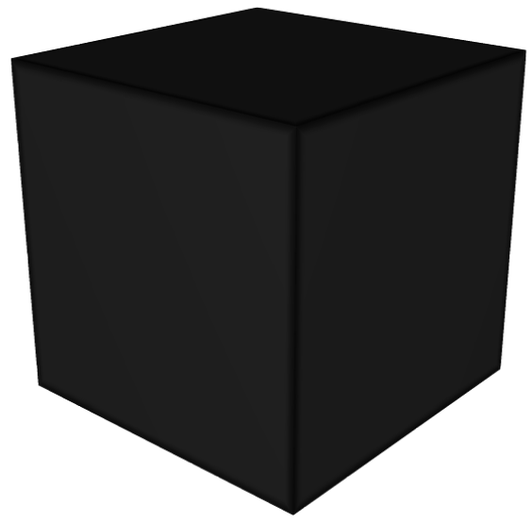
The target image

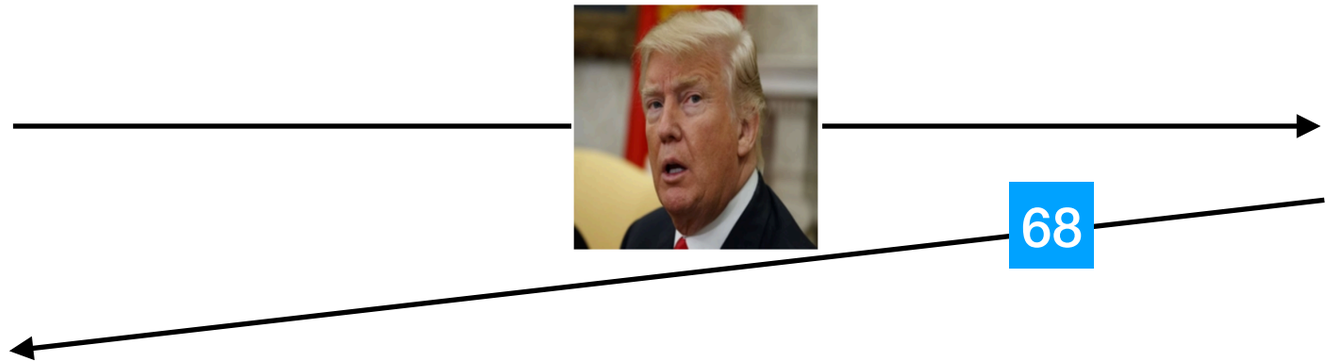
5

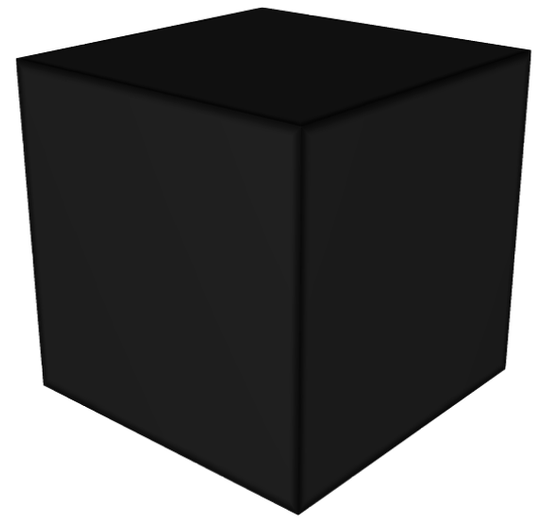


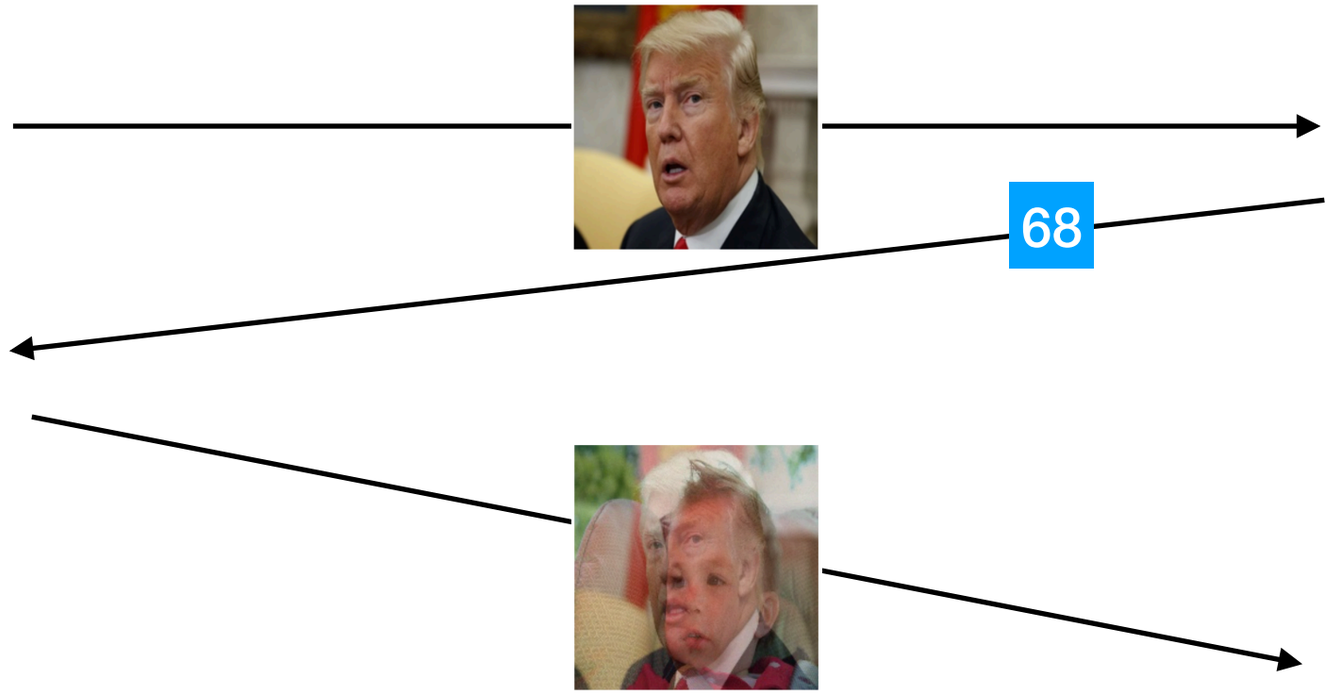


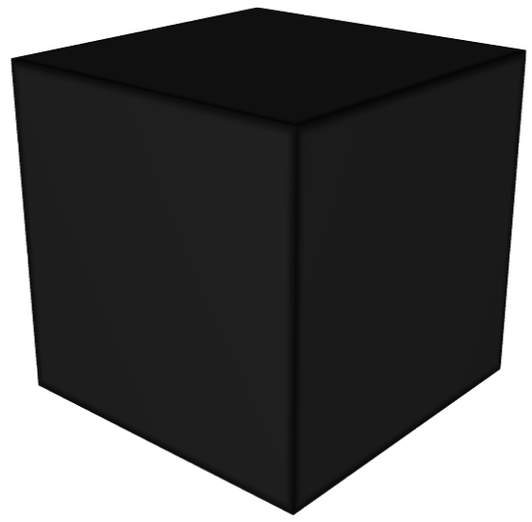


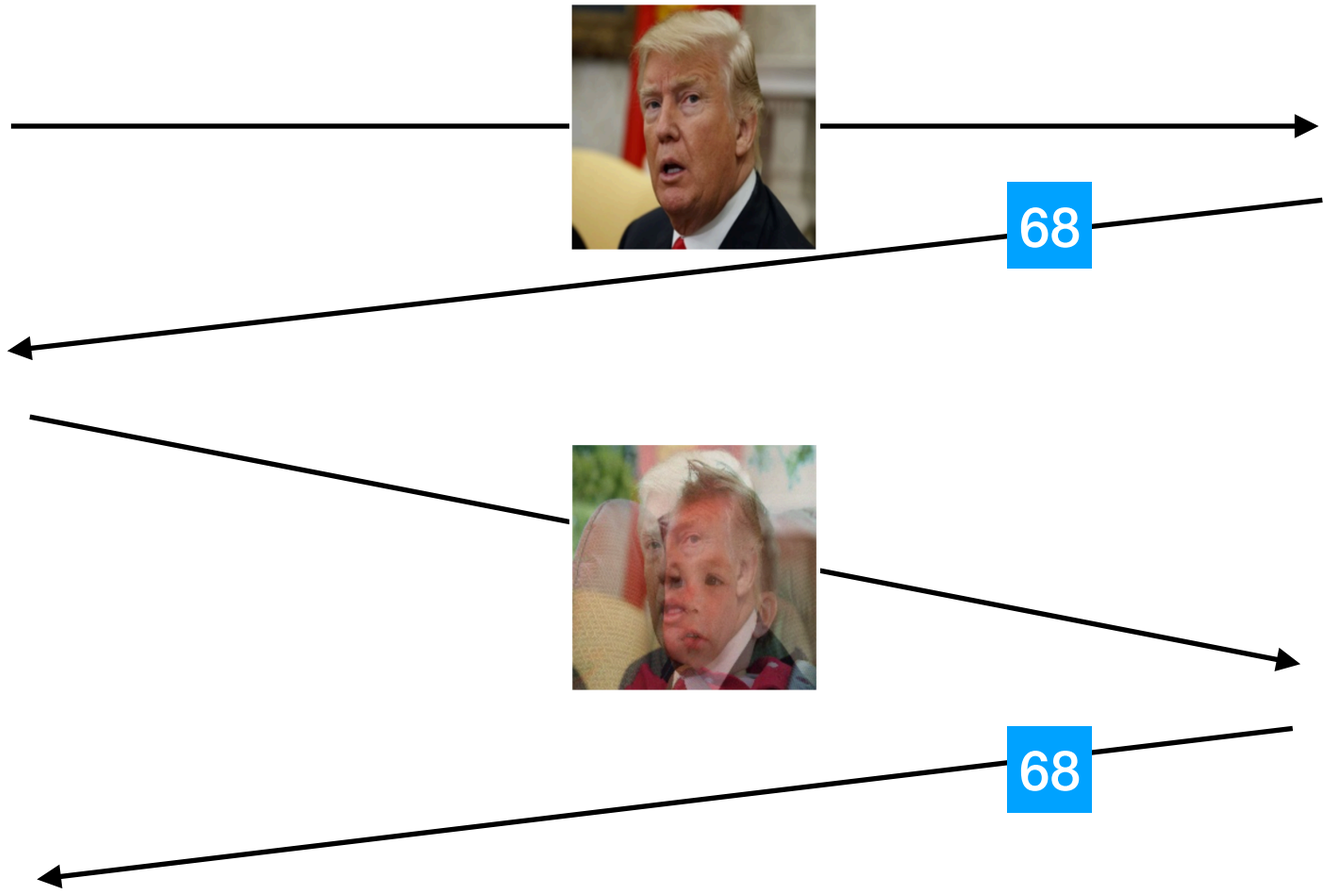


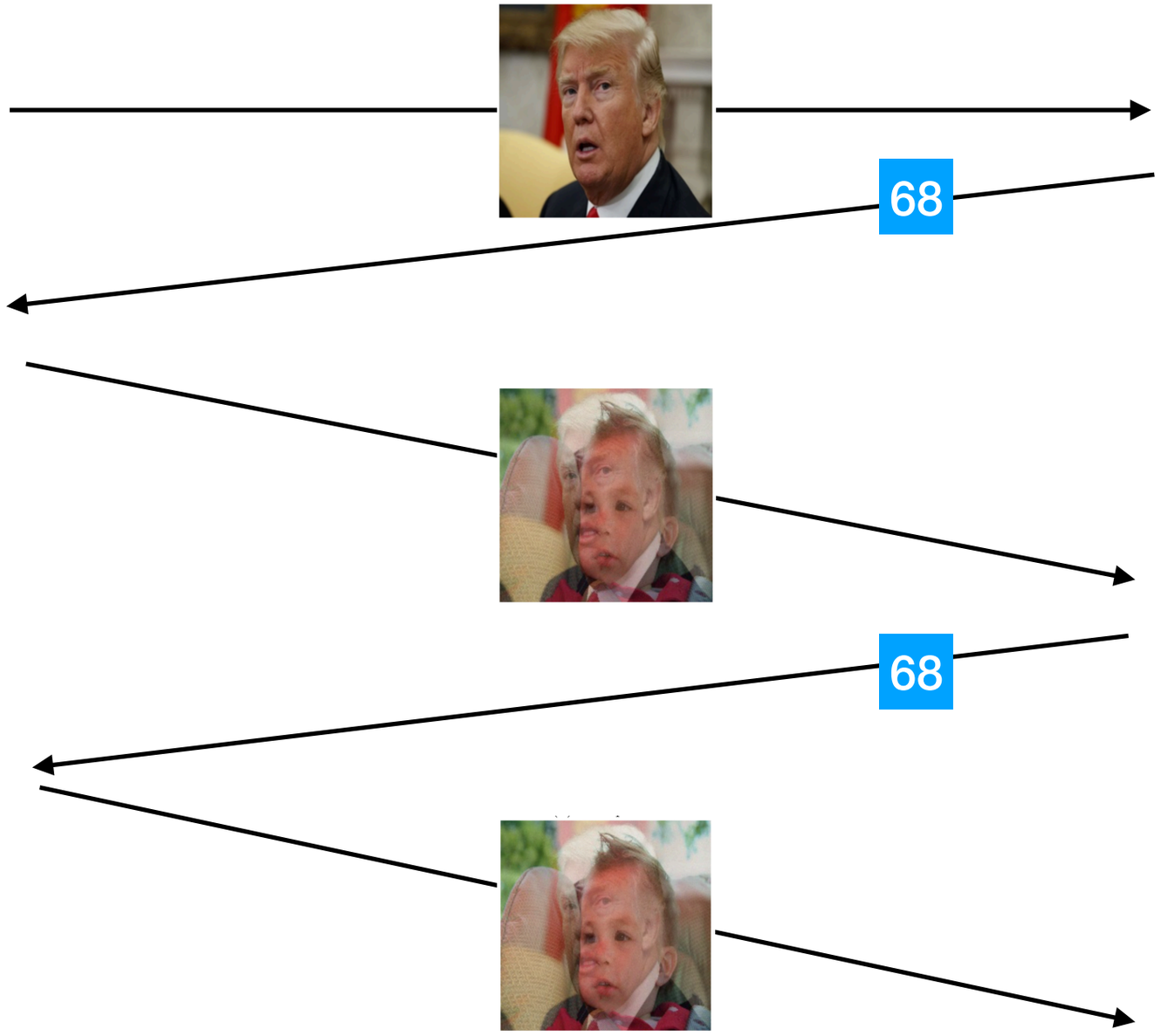

68




68



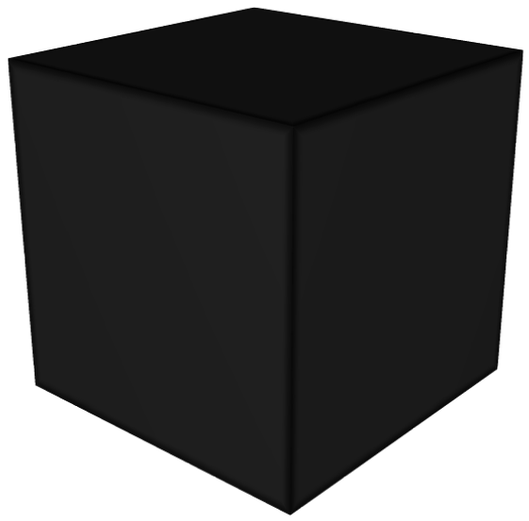

68
68

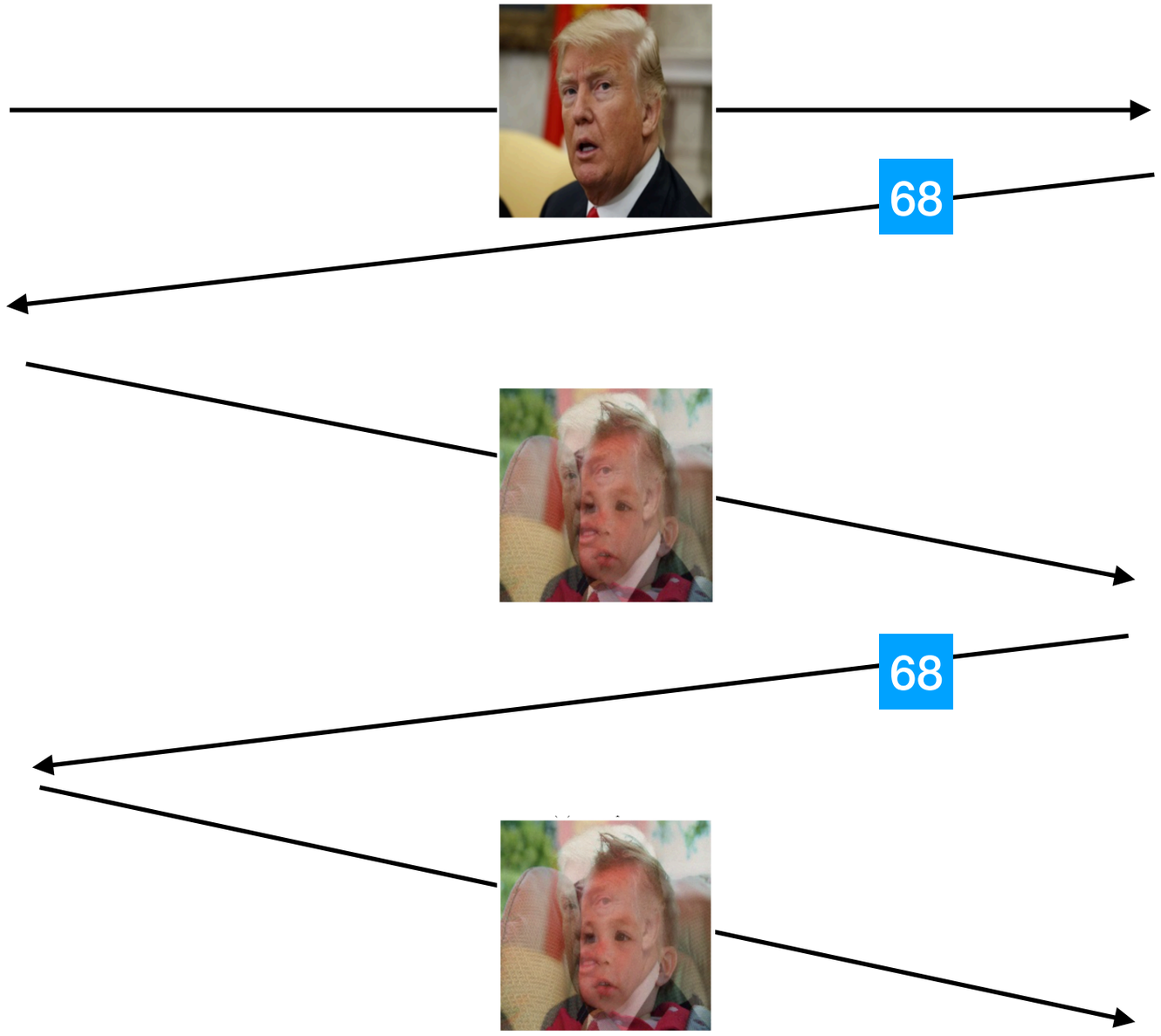

68
68

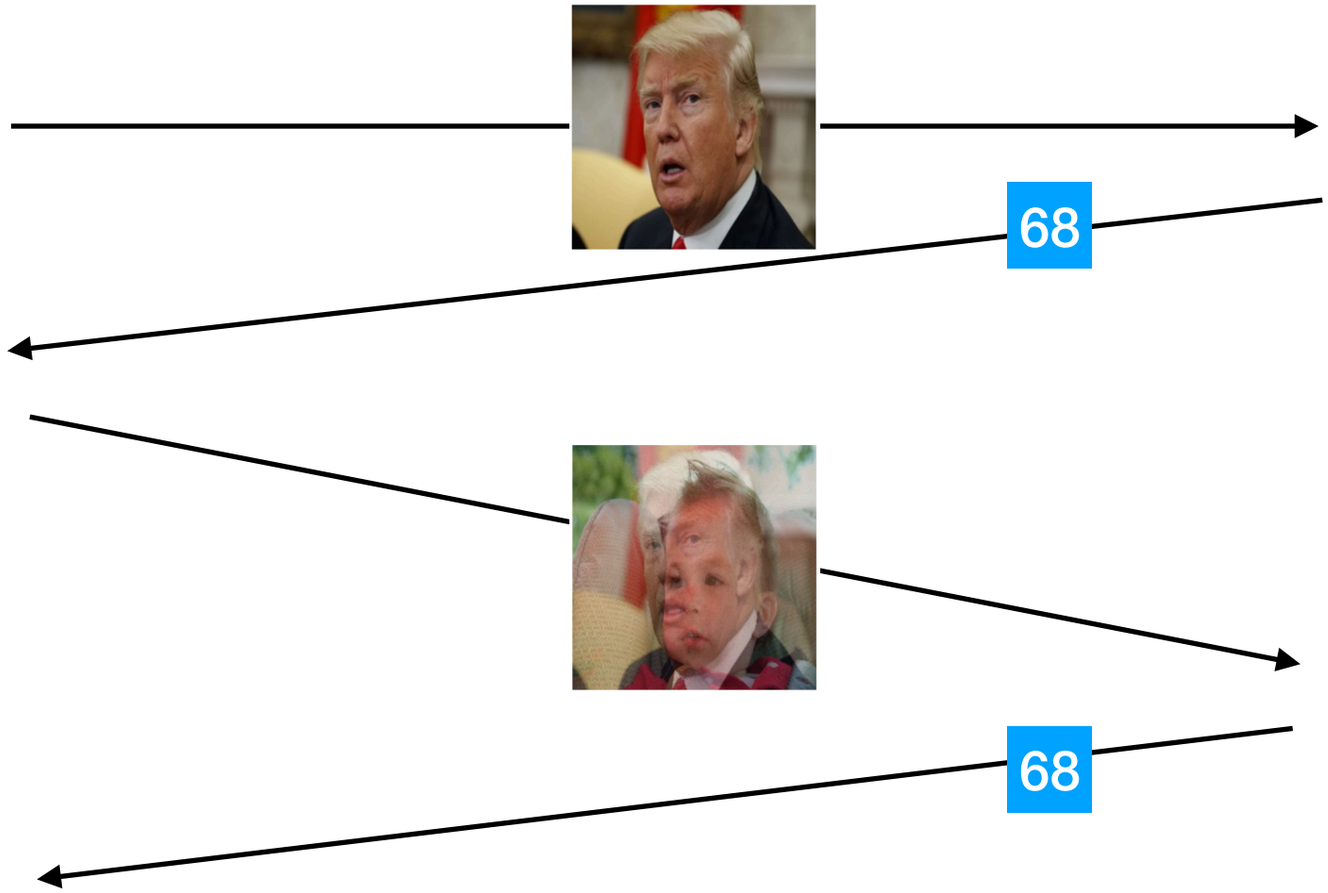
68

68

7

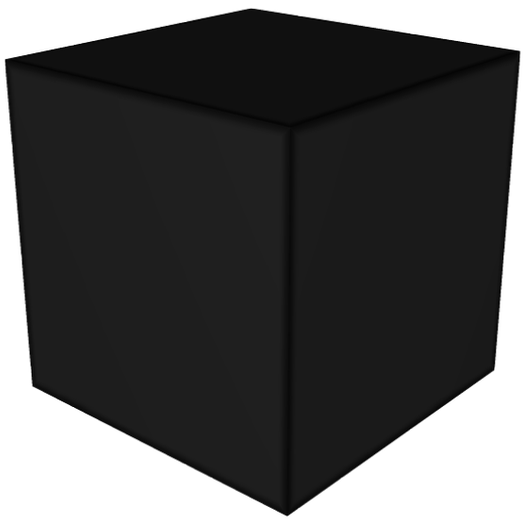



68
68

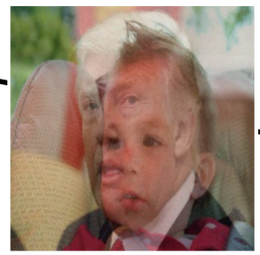

68
68

68

68

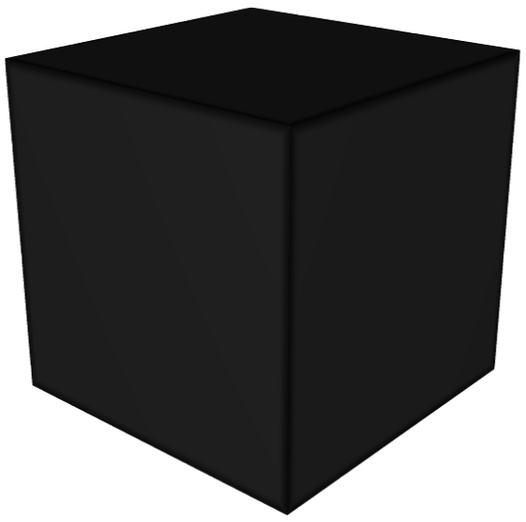


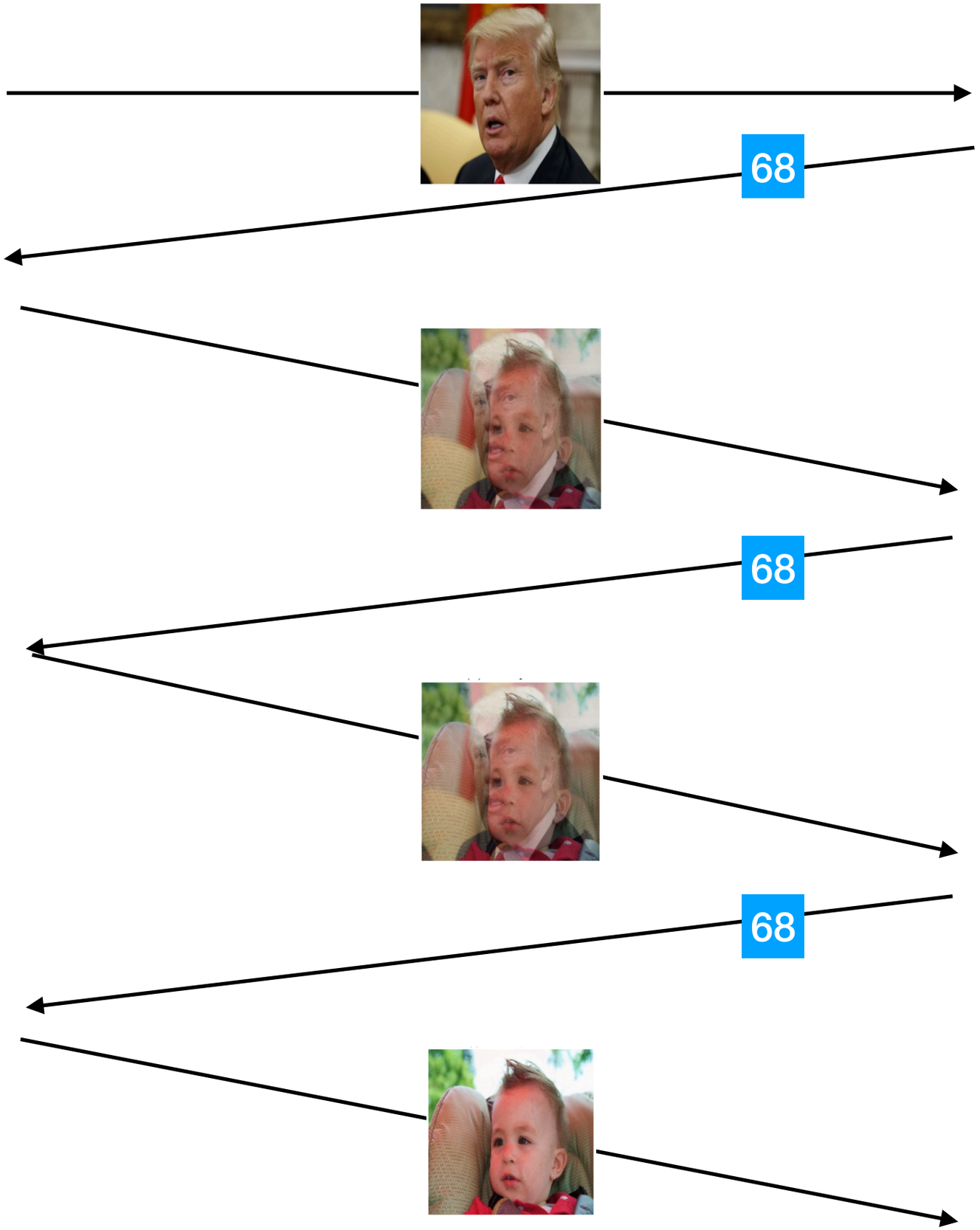
68

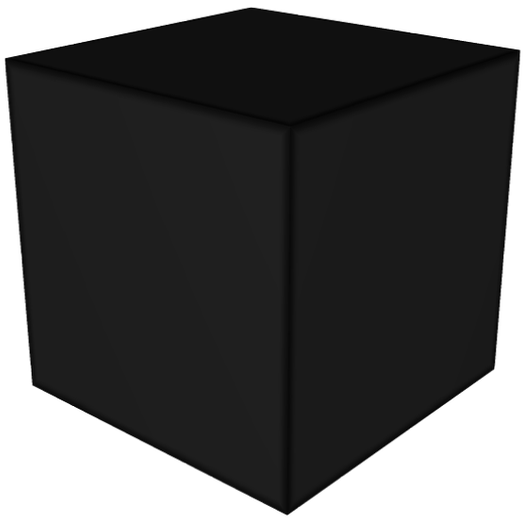


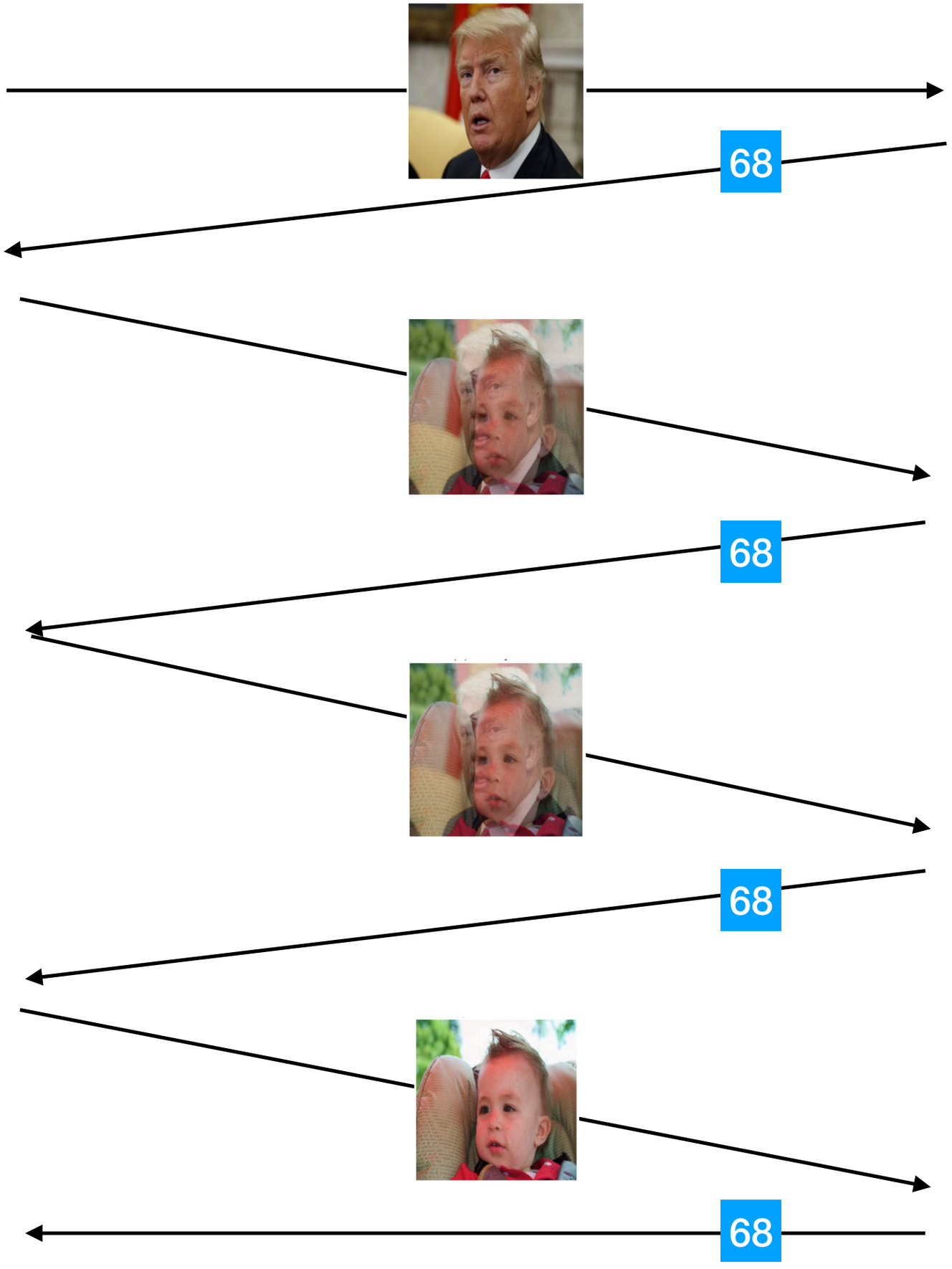
68

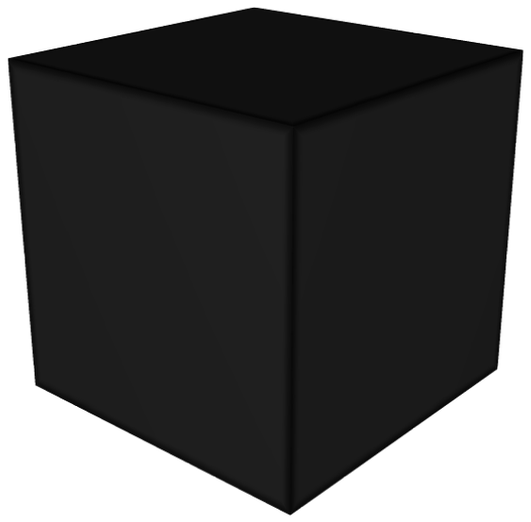
68



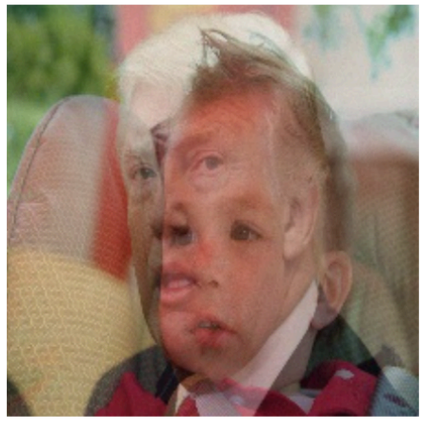

68
68
68




68
68
68
68



(a) The starting image



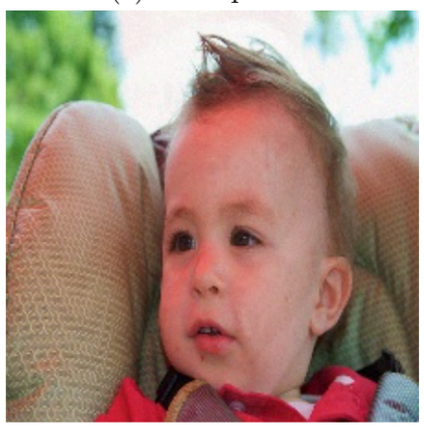
(b) 1000 queries

(c) 2000 queries

(d) 4000 queries

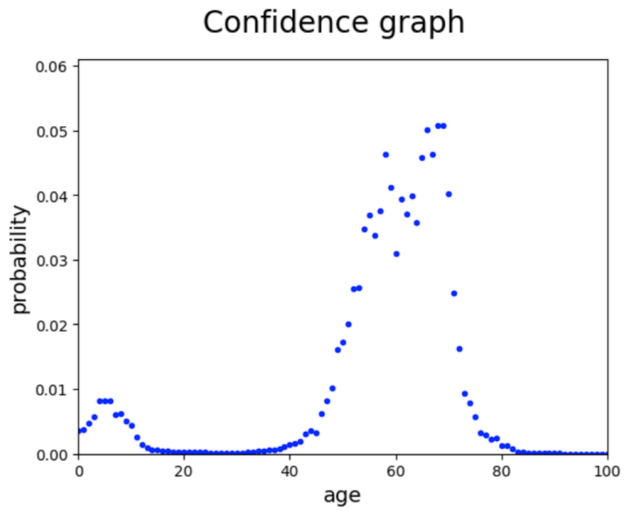
(e) 8000 queries

(f) 12000 queries

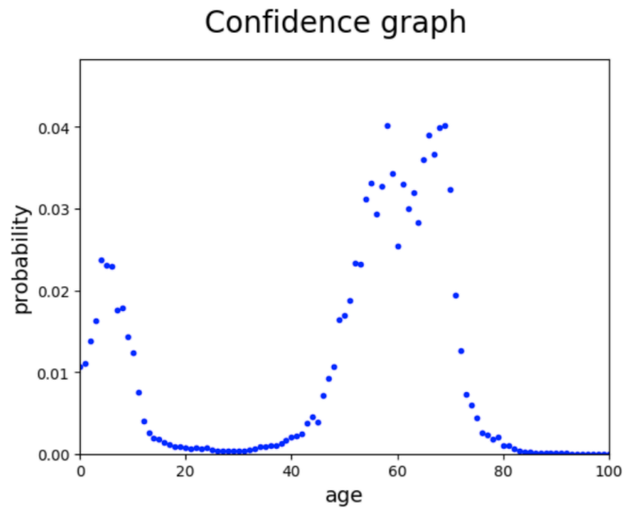


(g) 16000 queries

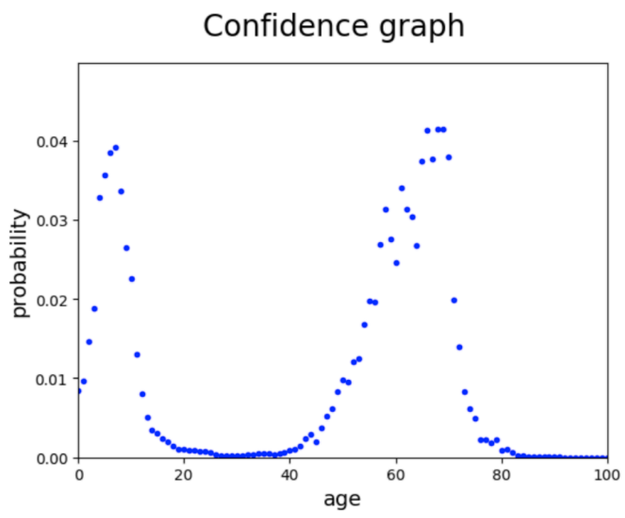
(h) Final adversarial sample



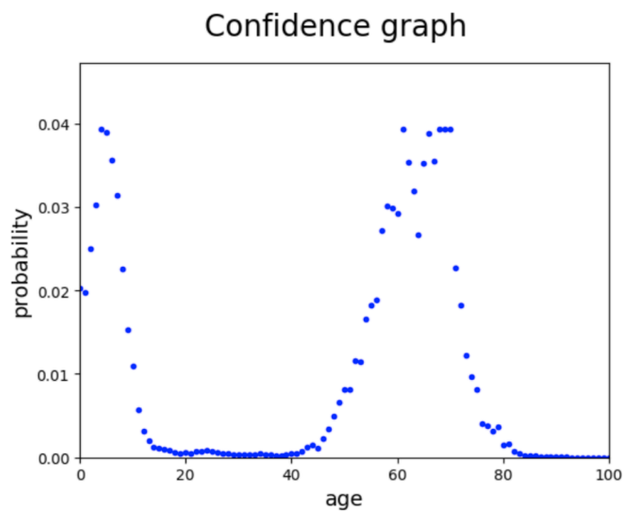
(a) 3000 queries



(b) 4000 queries

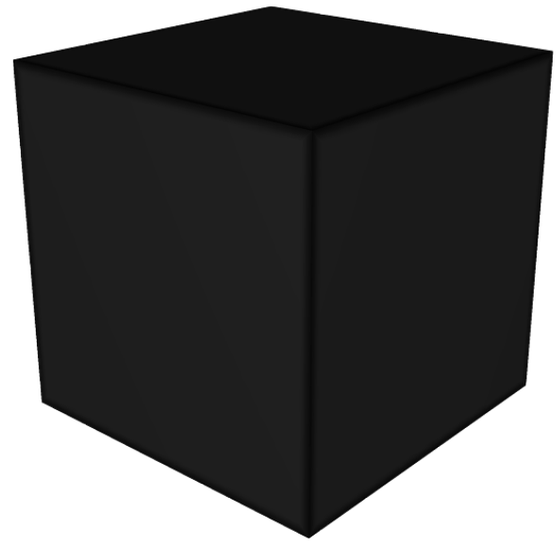


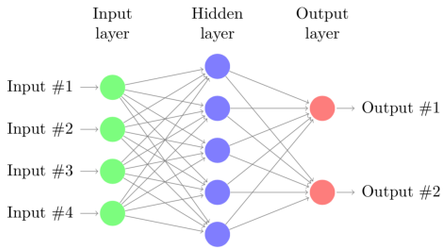
(c) 8000 queries

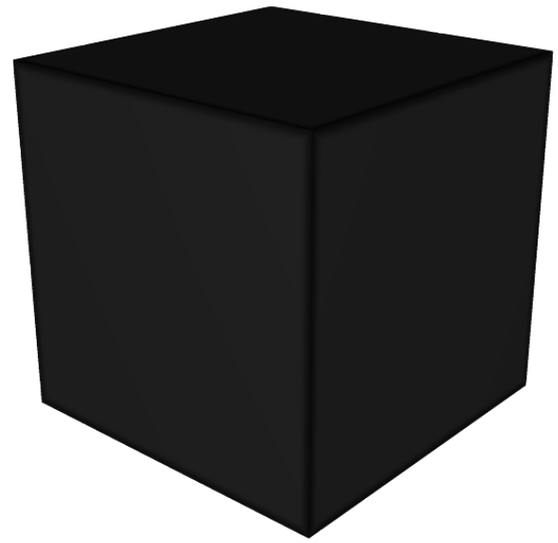


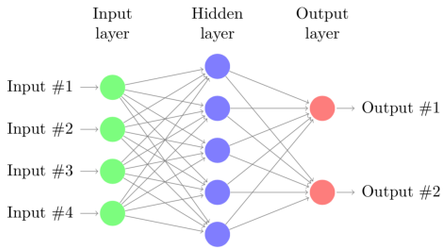
(d) The final adversarial sample (77 000 queries)

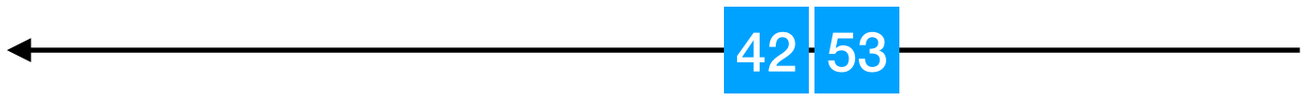
Transfer Based Approach

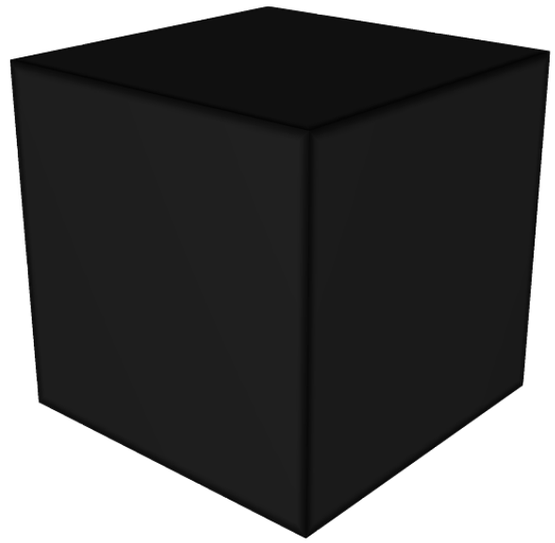


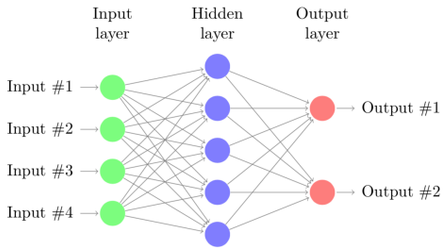

Input layer
Hidden layer
Output layer
Input #1
Input #2
Input #3
Input #4
Output #1
Output #2




Input layer
Hidden layer
Output layer
Input #1
Input #2
Input #3
Input #4
Output #1
Output #2

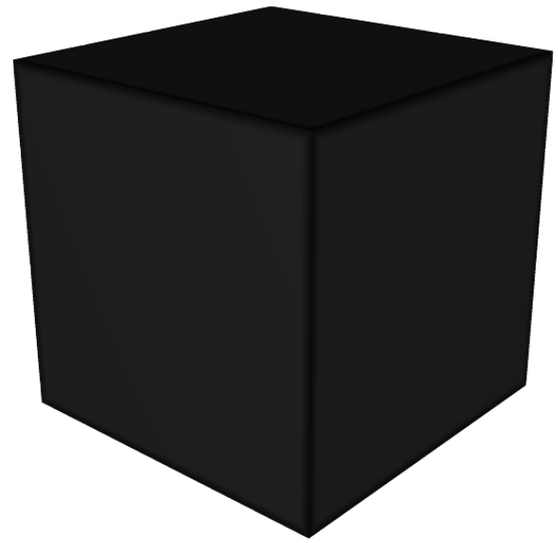

42
53

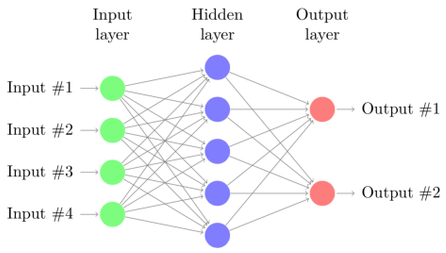


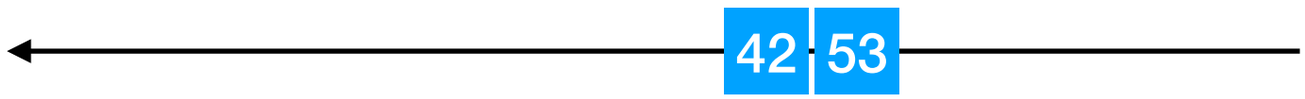


42 53

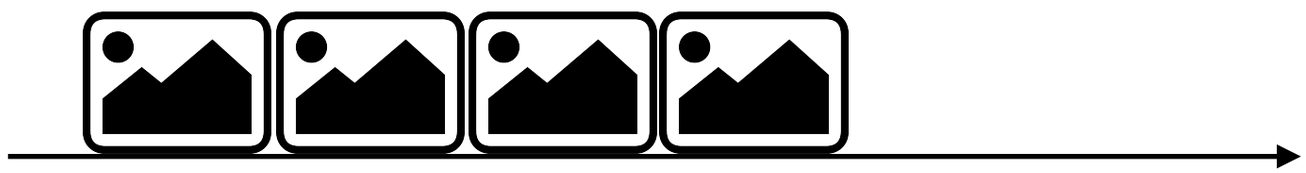
Train NN

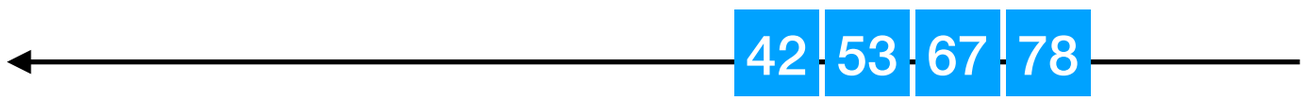


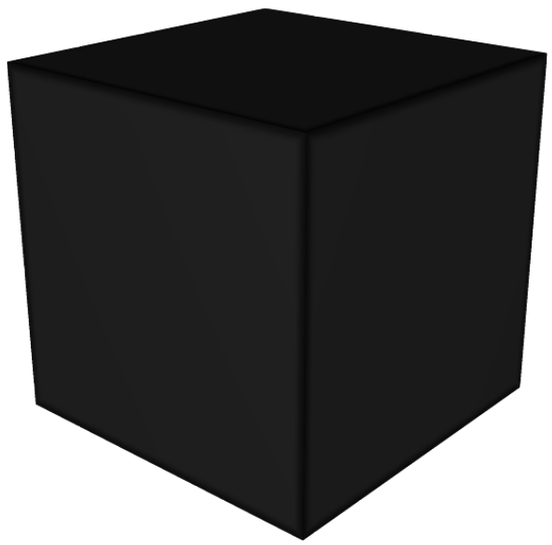

Input layer
Hidden layer
Output layer
Input #1
Input #2
Input #3
Input #4
Output #1
Output #2

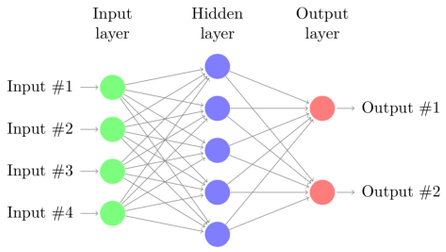

42
53

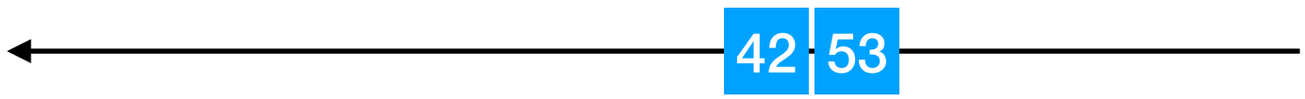
Train NN



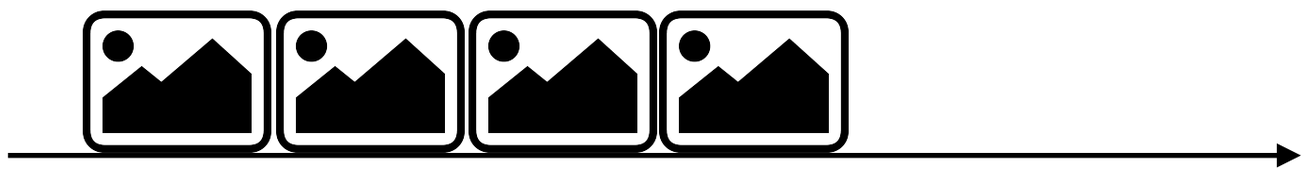

42
53
67
78

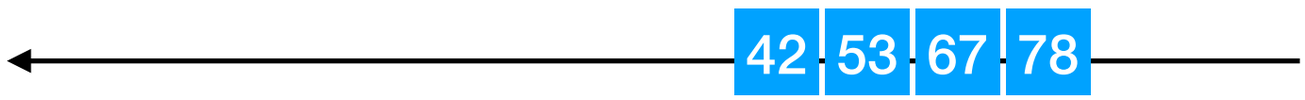


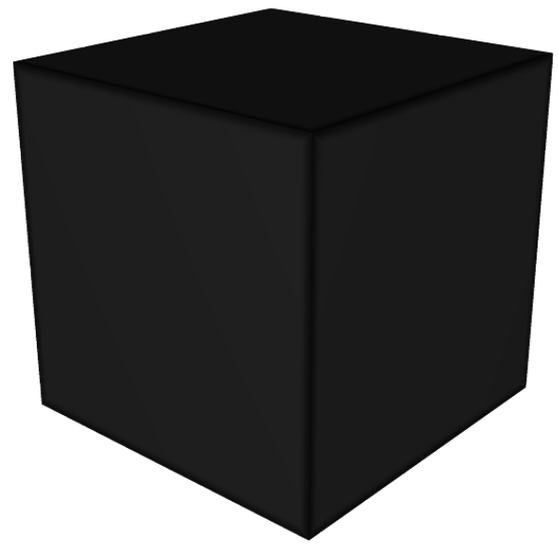




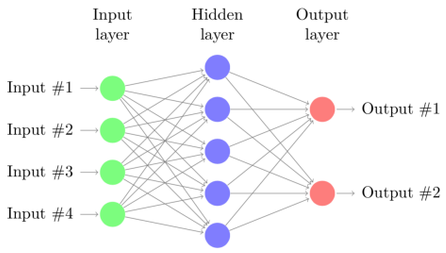
Train NN

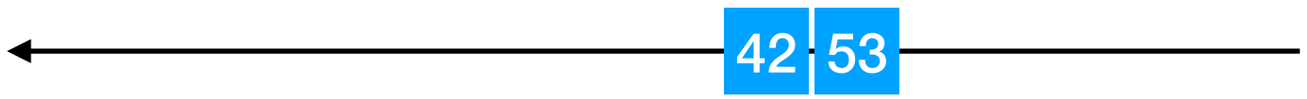




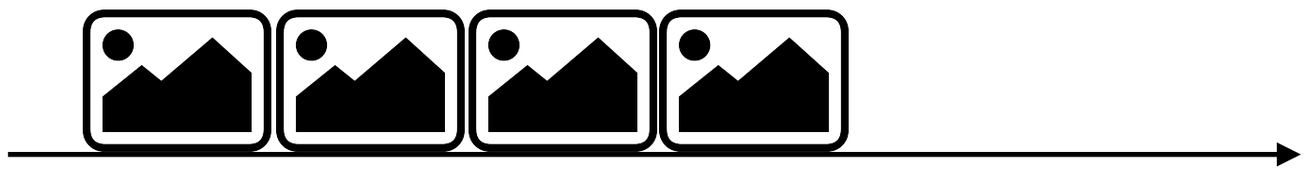


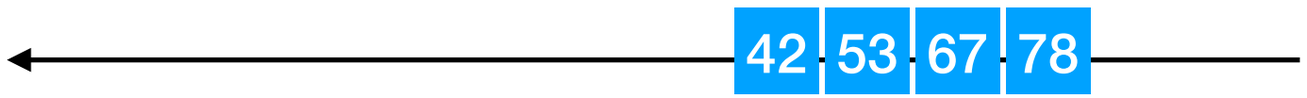
Train NN

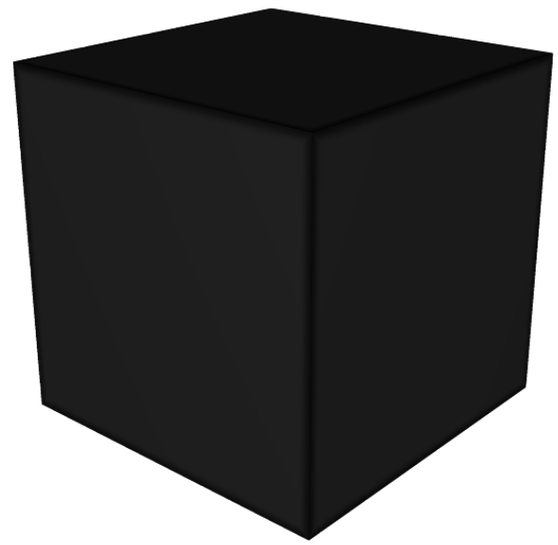




Train NN

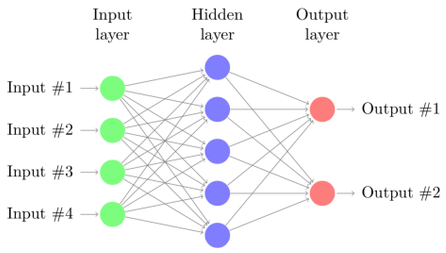


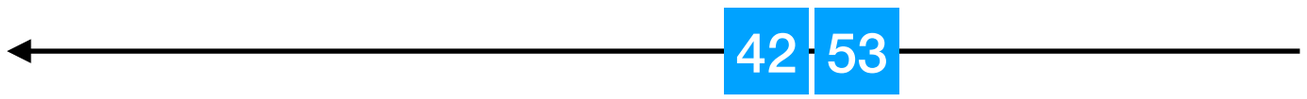




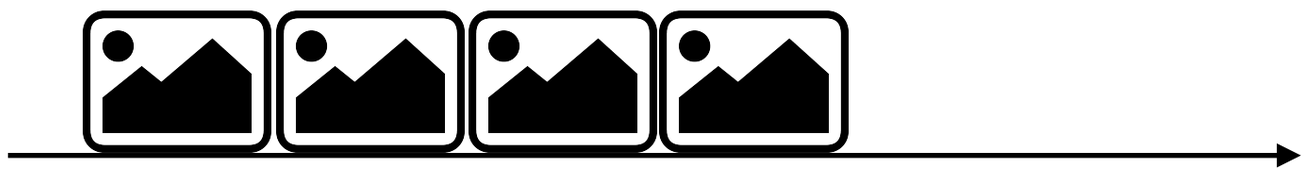
Train NN

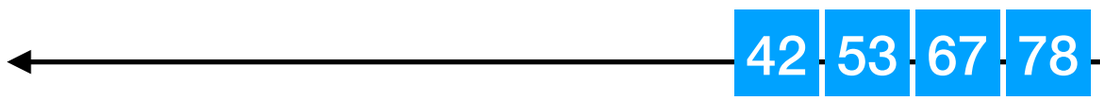
Attack white-box NN: FGSM or CW

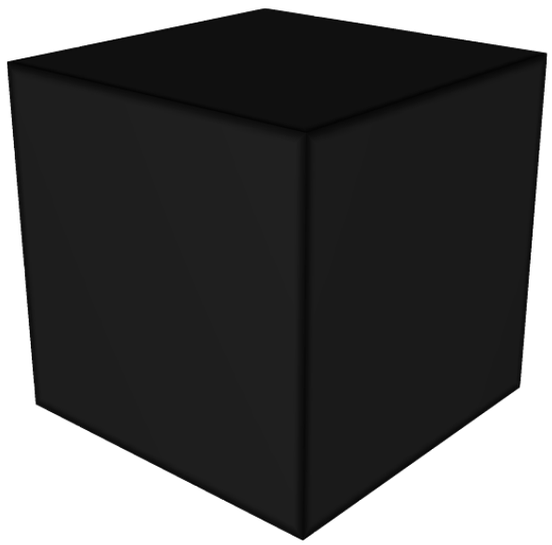




Train NN

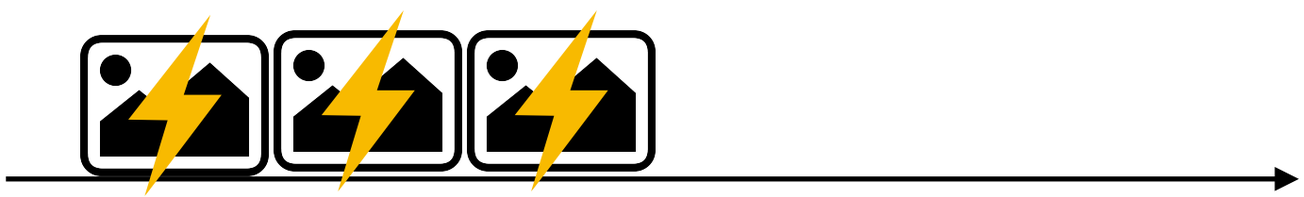


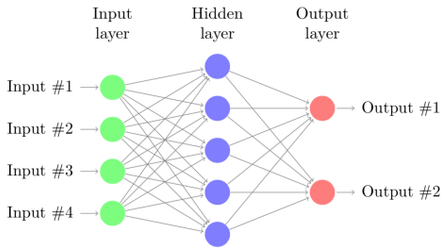


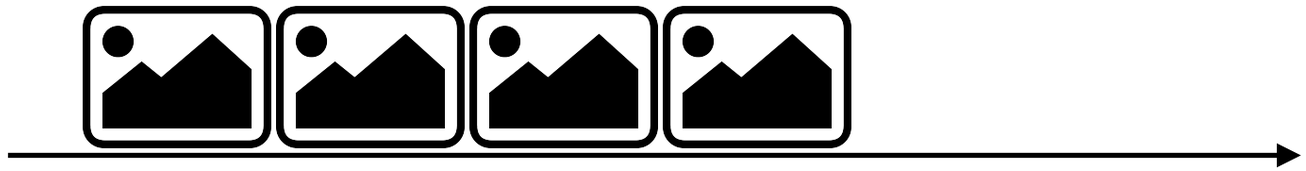


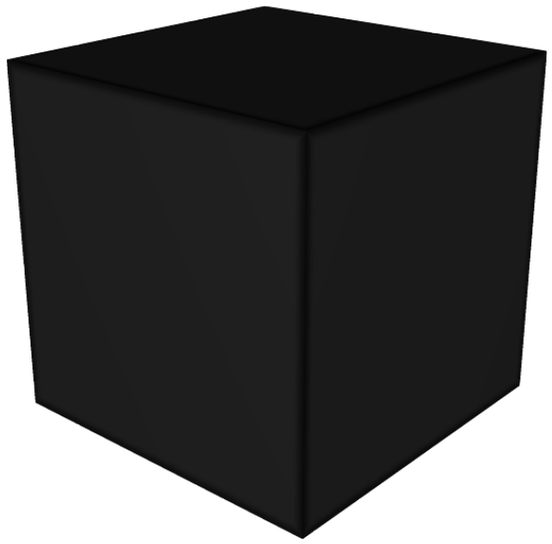
Train NN

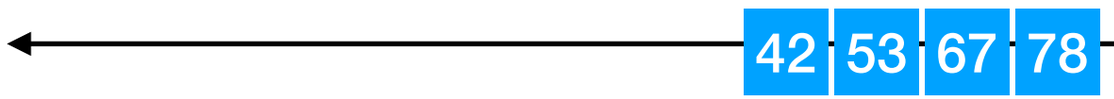
Attack white-box NN: FGSM or CW





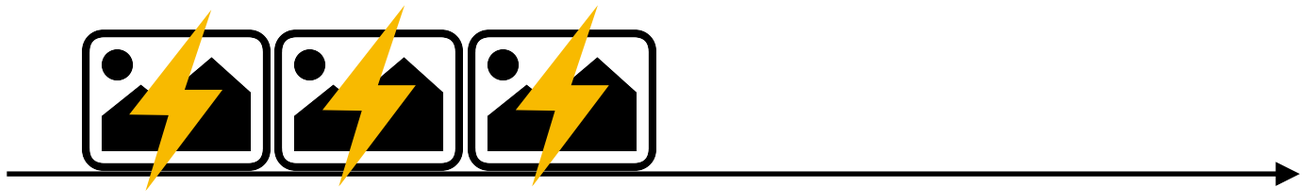


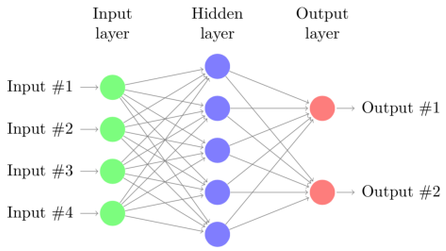


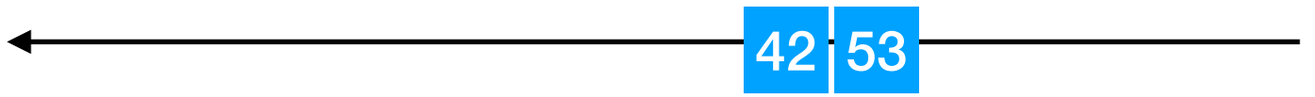


Train NN

Attack white-box NN: FGSM or CW

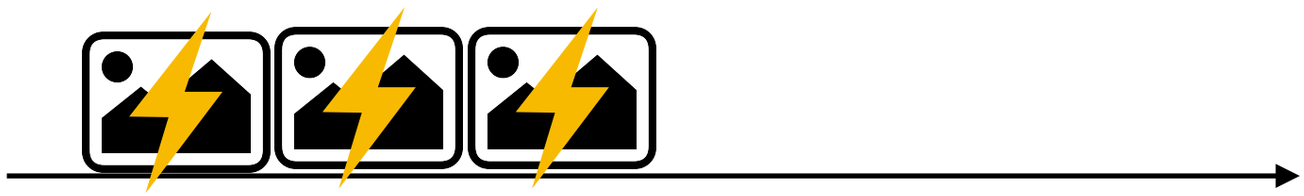


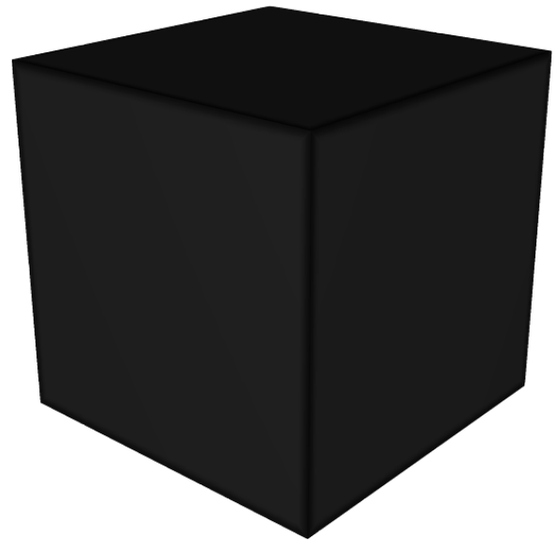




Train NN

Attack white-box NN: FGSM or CW





1. High memory expectation for the dataset expansion

2. Poor results without the dataset expansion

1. High memory expectation for the dataset expansion

2. Poor results without the dataset expansion

3. Misclassification transfers better than targeted misclassification*

* Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song
Delving into transferable adversarial examples and black-box attacks
CoRR, abs/1611.02770, 2016.

1. High memory expectation for the dataset expansion

2. Poor results without the dataset expansion

3. Misclassification transfers better than targeted misclassification*

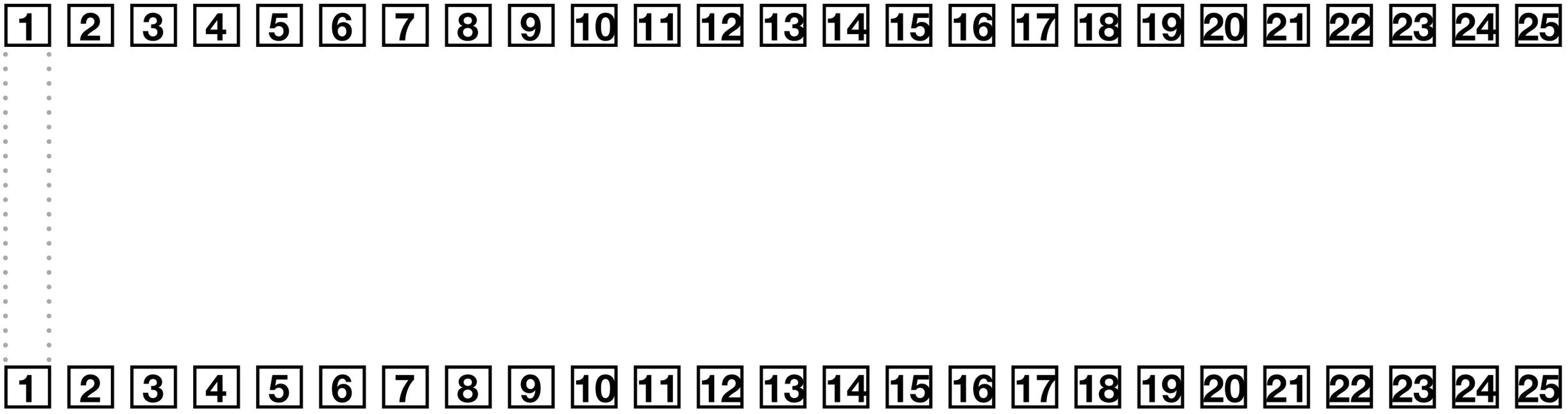
The Semi-targeted Approach

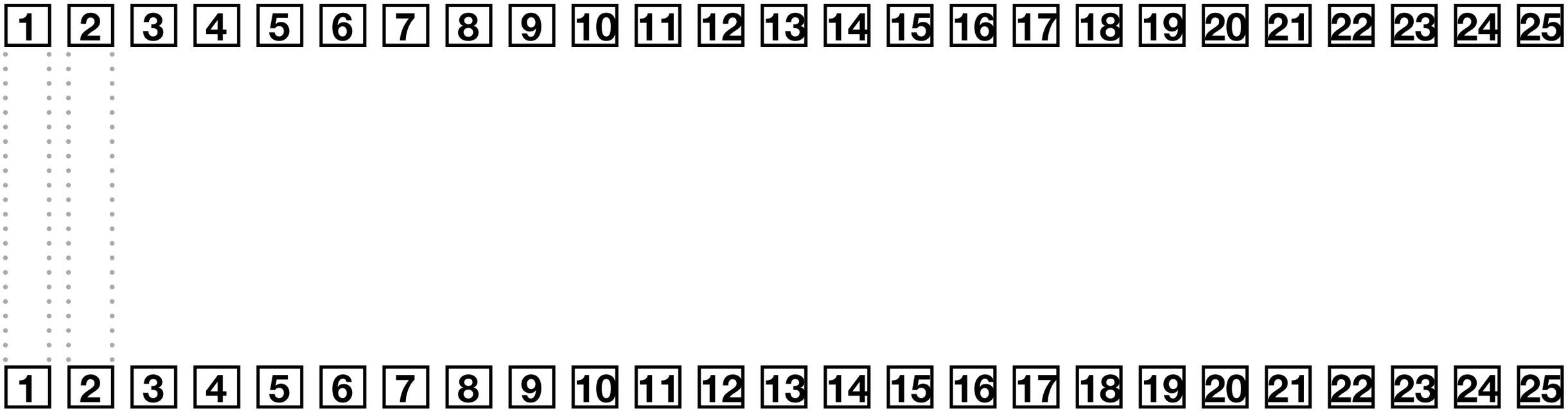
* Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song
Delving into transferable adversarial examples and black-box attacks
CoRR, abs/1611.02770, 2016.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

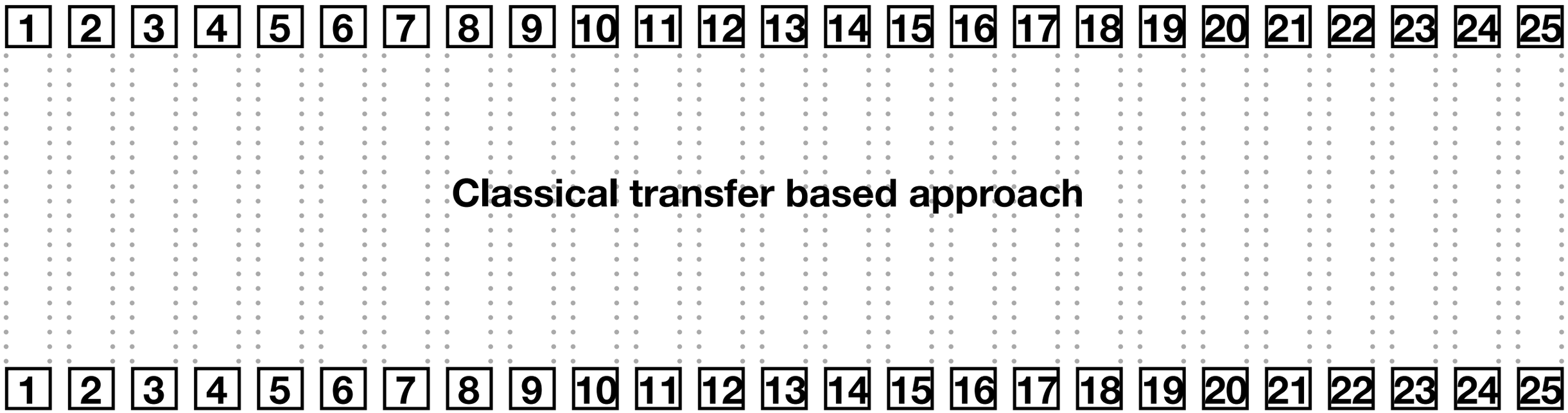
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

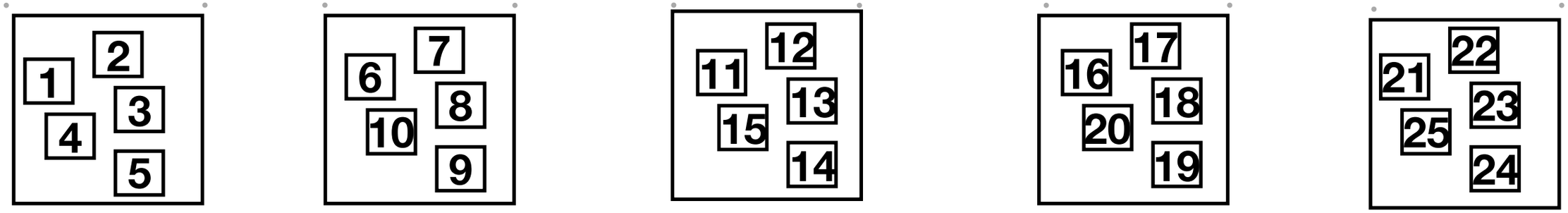

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Classical transfer based approach
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Classical transfer based approach

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

The Semi-targeted approach



The semi-targeted approach: results

- 4 "buckets": 0-25, 26-50, 51-75, 76-100
- FGSM
- **Managed to execute 3 iterations of dataset expansion**

	clean samples MAE	transfer based approach MAE	semi-targeted approach MAE
model 1	6.08	7.57	6.00
model 2	6.99	10.66	9.25
model 3	5.29	**5.57**	**7.75**
model 4	3.75	**4.25**	**4.83**

Contributions

- A framework for the white-box and the black-box attacks is developed
- Adversarial algorithms are evaluated in different environments
- The new semi-targeted black-box attack is introduced

Contributions

- A framework for the white-box and the black-box attacks is developed

- Adversarial algorithms are evaluated in different environments

- The new semi-targeted black-box attack is introduced

Thank you!

Appendix A

CW

Carlini and Wagner (CW)

Carlini and Wagner (CW)

1. *Minimize* added perturbation
2. Must be adversarial (target class)
3. Must be a valid image

Carlini and Wagner (CW)

1. *Minimize* added perturbation

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } \mathcal{C}(\boldsymbol{x}+\boldsymbol{\delta}) = t \\ &\boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

2. Must be adversarial (target class)

3. Must be a valid image

Carlini and Wagner (CW)

1. *Minimize* added perturbation

2. Must be adversarial (target class)

3. Must be a valid image

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } \mathcal{C}(\boldsymbol{x}+\boldsymbol{\delta}) = t \\ &\qquad \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

iff

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } f(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0 \\ &\qquad \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

Carlini and Wagner (CW)

1. *Minimize* added perturbation

2. Must be adversarial (target class)

3. Must be a valid image

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } \mathcal{C}(\boldsymbol{x}+\boldsymbol{\delta}) = t \\ &\boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

↓

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } f(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0 \\ &\boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

iff

↓ $c > 0$

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) + c \cdot f(\boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

Carlini and Wagner (CW)

1. *Minimize* added perturbation

2. Must be adversarial (target class)

3. Must be a valid image

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } \mathcal{C}(\boldsymbol{x}+\boldsymbol{\delta}) = t \\ &\boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

$$\downarrow$$

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } f(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0 \\ &\boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

iff

$$\downarrow \quad c > 0$$

$$\begin{aligned} &\text{minimize } \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\delta}) + c \cdot f(\boldsymbol{x}+\boldsymbol{\delta}) \\ &\text{such that } \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^n \end{aligned}$$

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = max(max(\{\boldsymbol{Z}(\boldsymbol{x}+\boldsymbol{\delta})_i : i \neq t\}) - \boldsymbol{Z}(\boldsymbol{x}+\boldsymbol{\delta})_t, 0)$$

Appendix B

dataset augmentation

$$\nabla \mathbf{F}(\mathbf{X}) = \left[\frac{\partial \mathbf{F}(\mathbf{X})}{\partial x_1}, \frac{\partial \mathbf{F}(\mathbf{X})}{\partial x_2} \right]$$

$$S_{\rho+1} = \{\vec{x} + \lambda \cdot \mathrm{sgn}(J_F[\tilde{O}(\vec{x})]) : \vec{x} \in S_\rho\} \cup S_\rho$$

Appendix C

Detailed results

The semi-targeted approach: results

- 4 "buckets": 0-25, 26-50, 51-75, 76-100
- FGSM
- **Managed to execute 3 iterations of dataset expansion**

	clean samples		adversarial samples					
blackbox model id	blackbox MAE	substitute accuracy	blackbox MAE	substitute accuracy	avg L2	std dev L2	min L2	max L2
1	6.08	0.67	6.00	0.37	2525.36	76.90	2102.57	2589.41
2	6.99	0.68	9.25	0.30	2525.20	76.96	2103.83	2589.41
3	5.29	0.69	7.75	0.35	2525.43	76.61	2108.77	2589.41
4	3.75	0.74	4.83	0.34	2525.48	76.57	2105.55	2589.41

Classical transfer based approach:

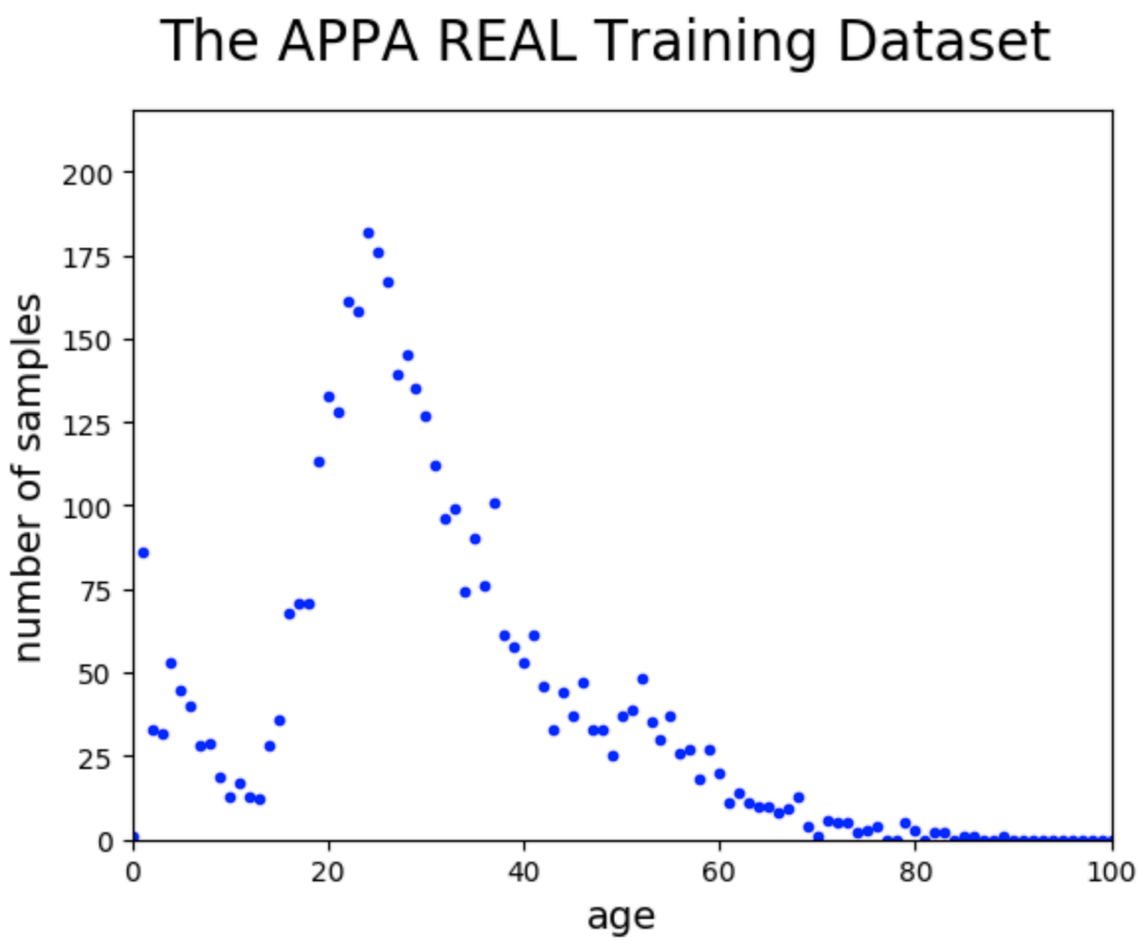
	clean samples		adversarial samples					
blackbox model id	blackbox MAE	substitute MAE	blackbox MAE	substitute MAE	avg L2	std dev L2	min L2	max L2
1	6.08	10.20	7.57	13.63	2524.84	77.62	2097.67	2589.41
2	6.99	11.09	**10.66**	18.59	2525.10	76.92	2105.15	2589.41
3	5.29	21.02	5.57	25.09	2527.32	73.56	2143.89	2589.41
4	3.75	8.25	4.25	14.75	2525.40	76.66	2111.09	2589.41

Appendix D

Dataset

The APPA-REAL dataset:

- total 7,591 images
  - 4113 training images
  - 1500 validation images
  - 1978 test images



The UTK-Face dataset:

- total 23,252 images