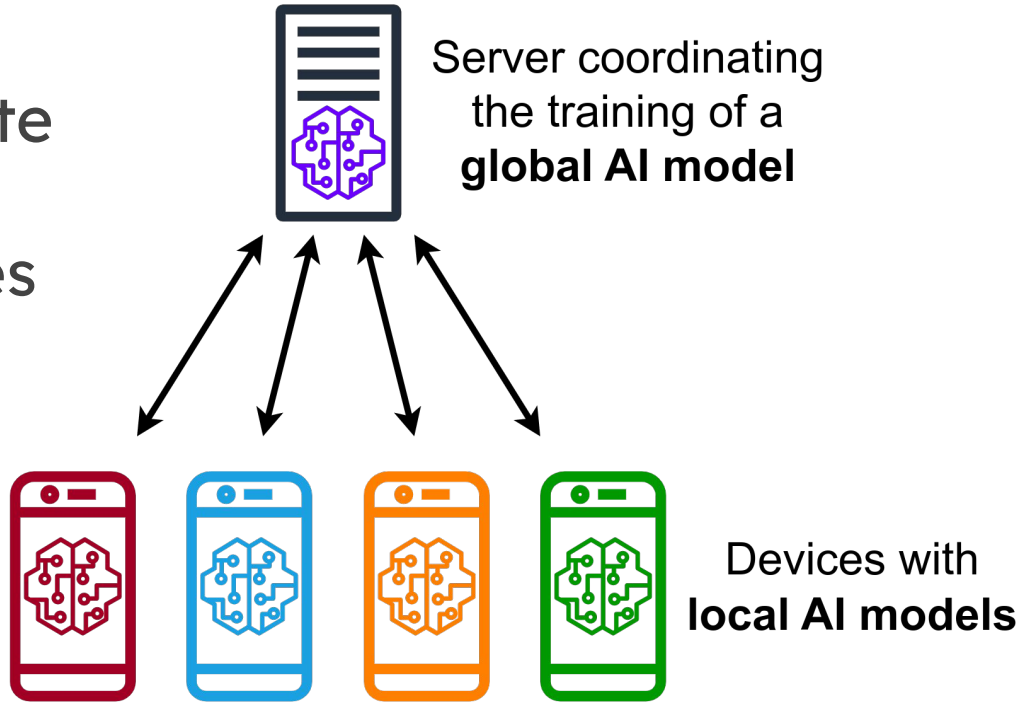
Federated Learning under Attack: Improving Gradient Inversion for Batch of Images

Luiz Leite, Yuri santos
Prof. Bruno Dalmazo,
Orientador: Prof. André Riker

UFPA
FURG

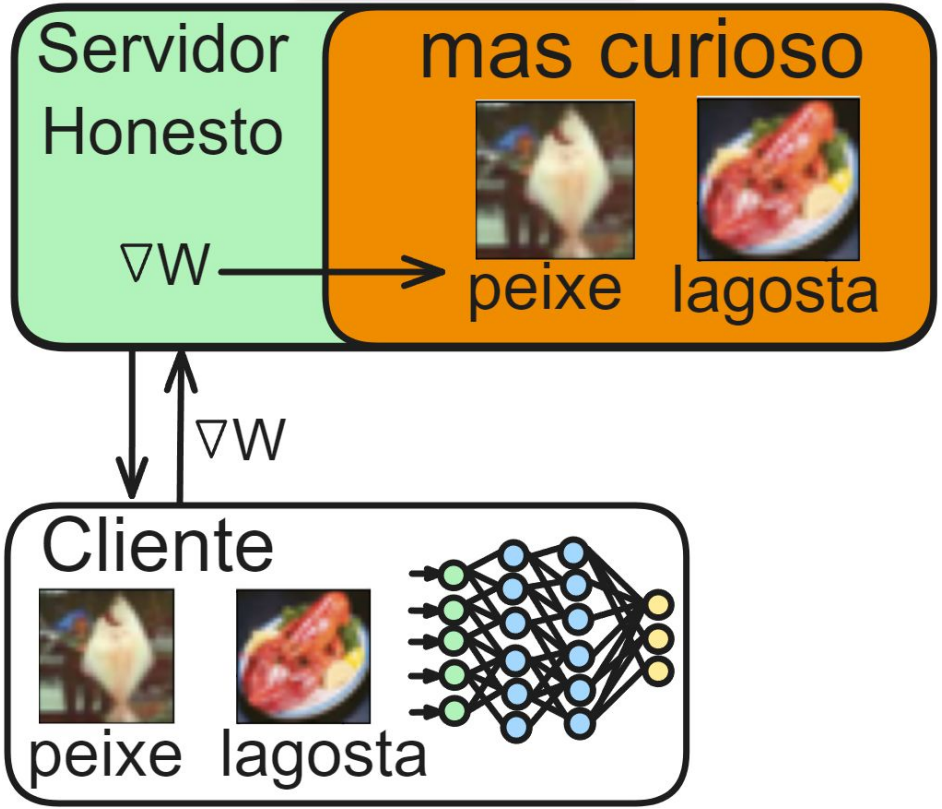
Motivação

- Tecnologia emergente
- Ameaças emergentes
- Dados sensíveis



Problema(s)

- Ameaças emergentes:
  - Inversão de gradiente
  - Servidor honesto, mas curioso
  - DLG e iDLG

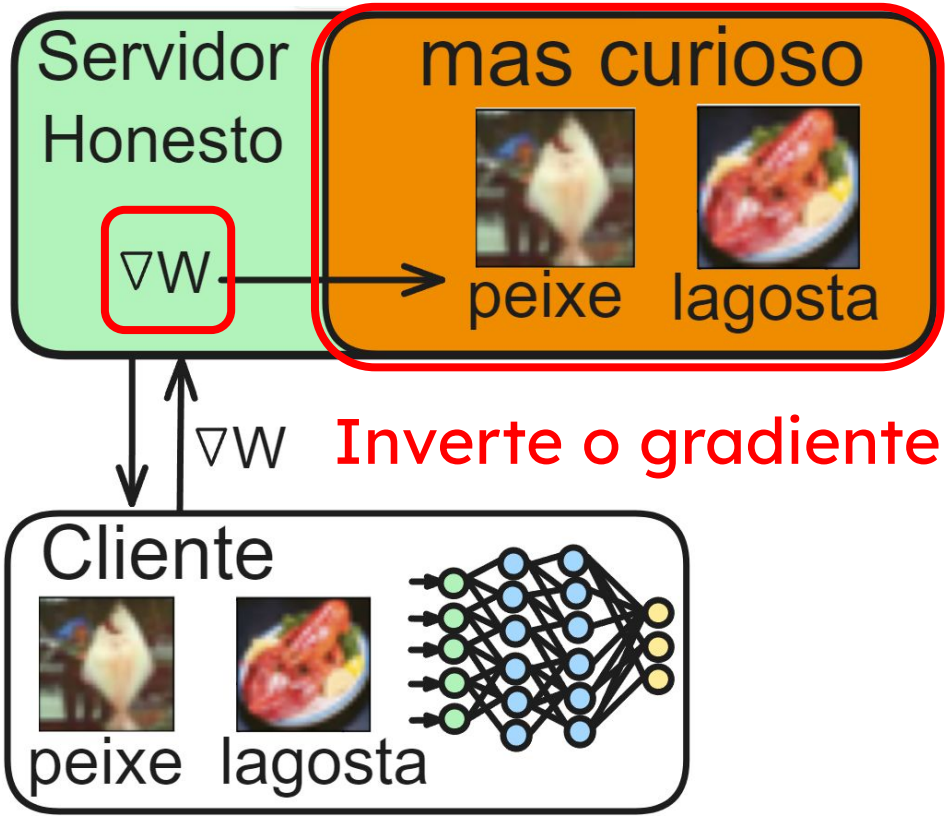


Problema(s)

- Ameaças emergentes:
  - Inversão de gradiente
  - Servidor honesto, mas curioso
  - DLG e iDLG

Problema(s)

- Ameaças emergentes:
  - Inversão de gradiente
  - Servidor honesto, mas curioso
  - DLG e iDLG

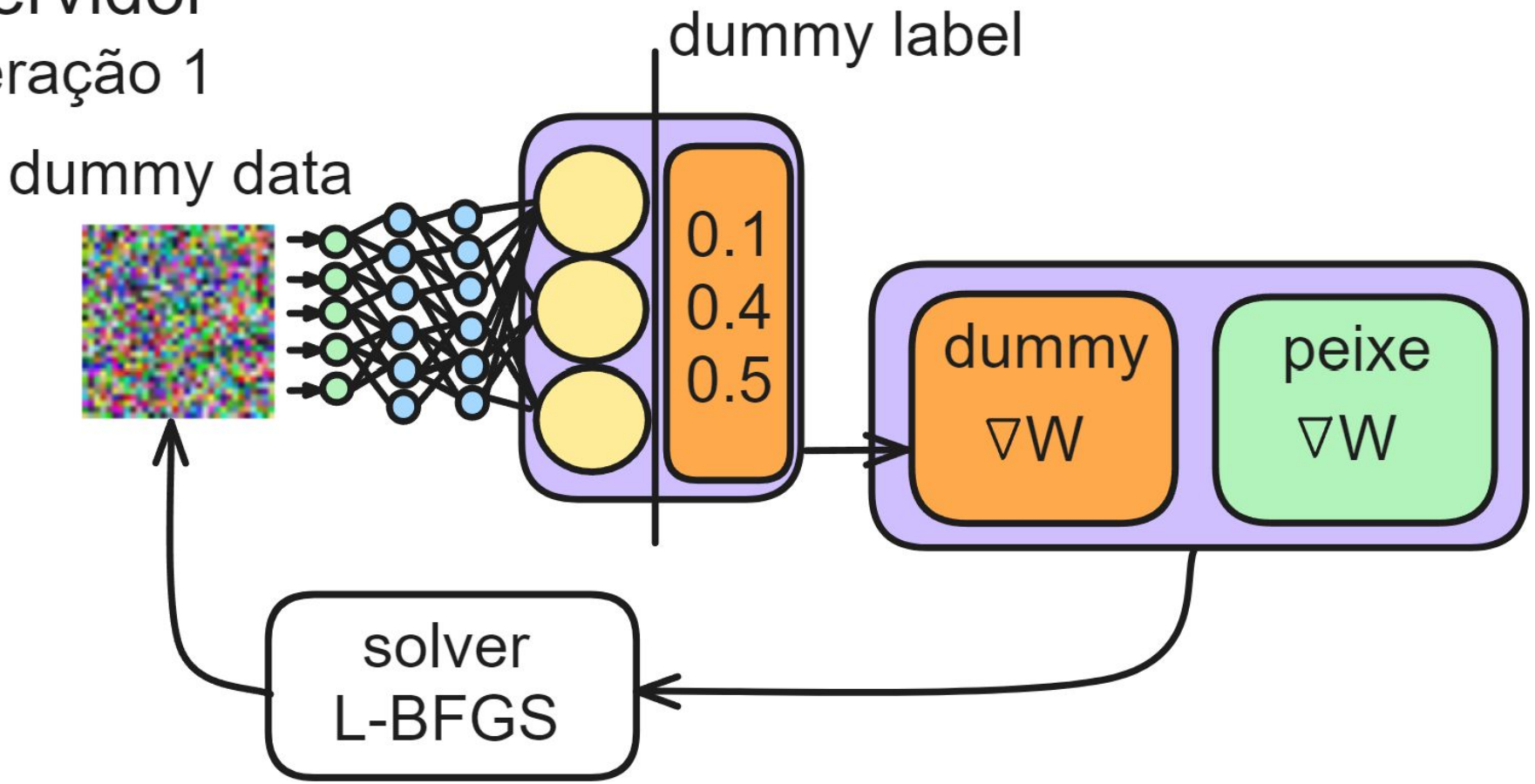


Problema(s)

- Ameaças emergentes:
  - Inversão de gradiente
  - Servidor honesto, mas curioso
  - DLG e iDLG

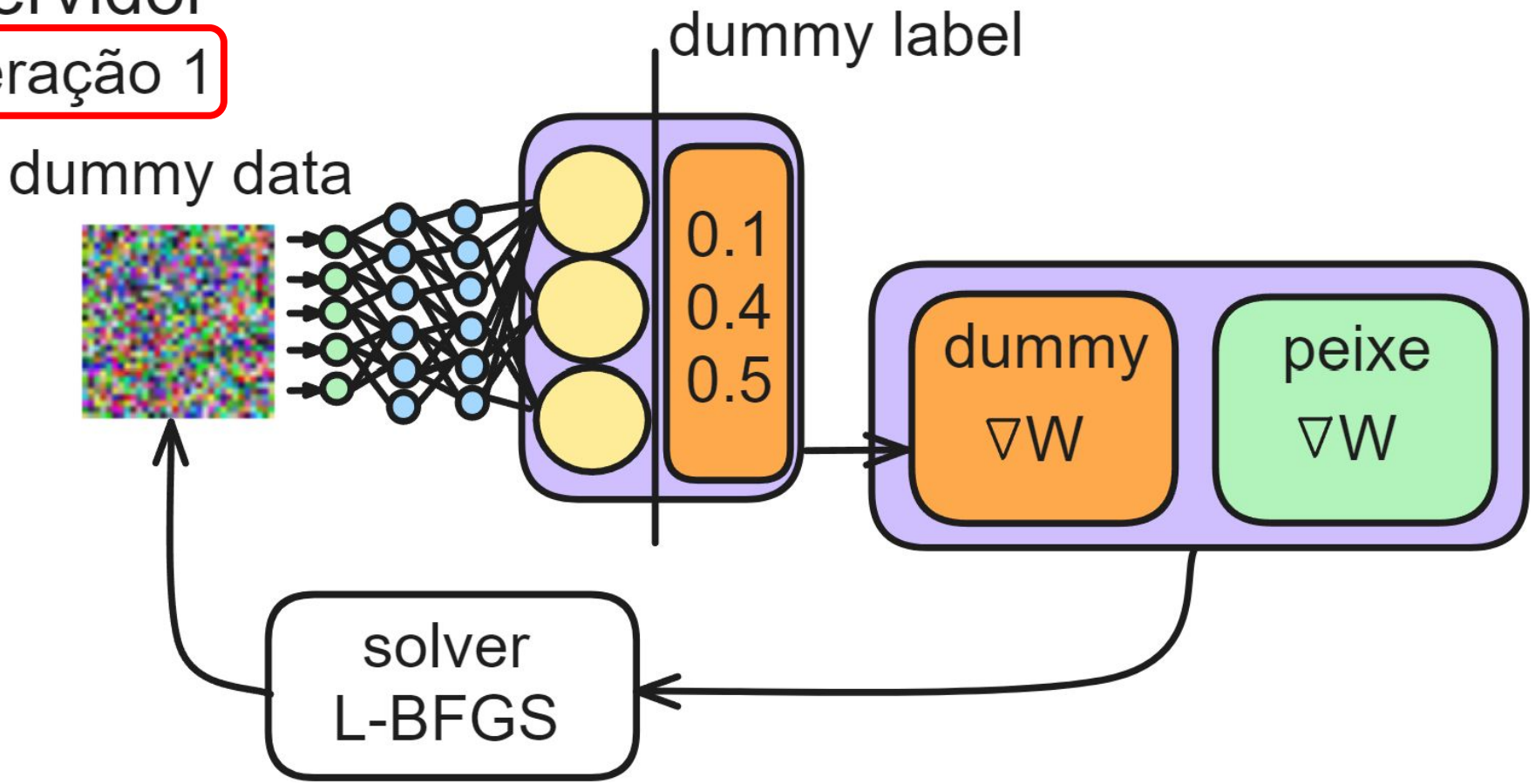
Servidor

Iteração 1



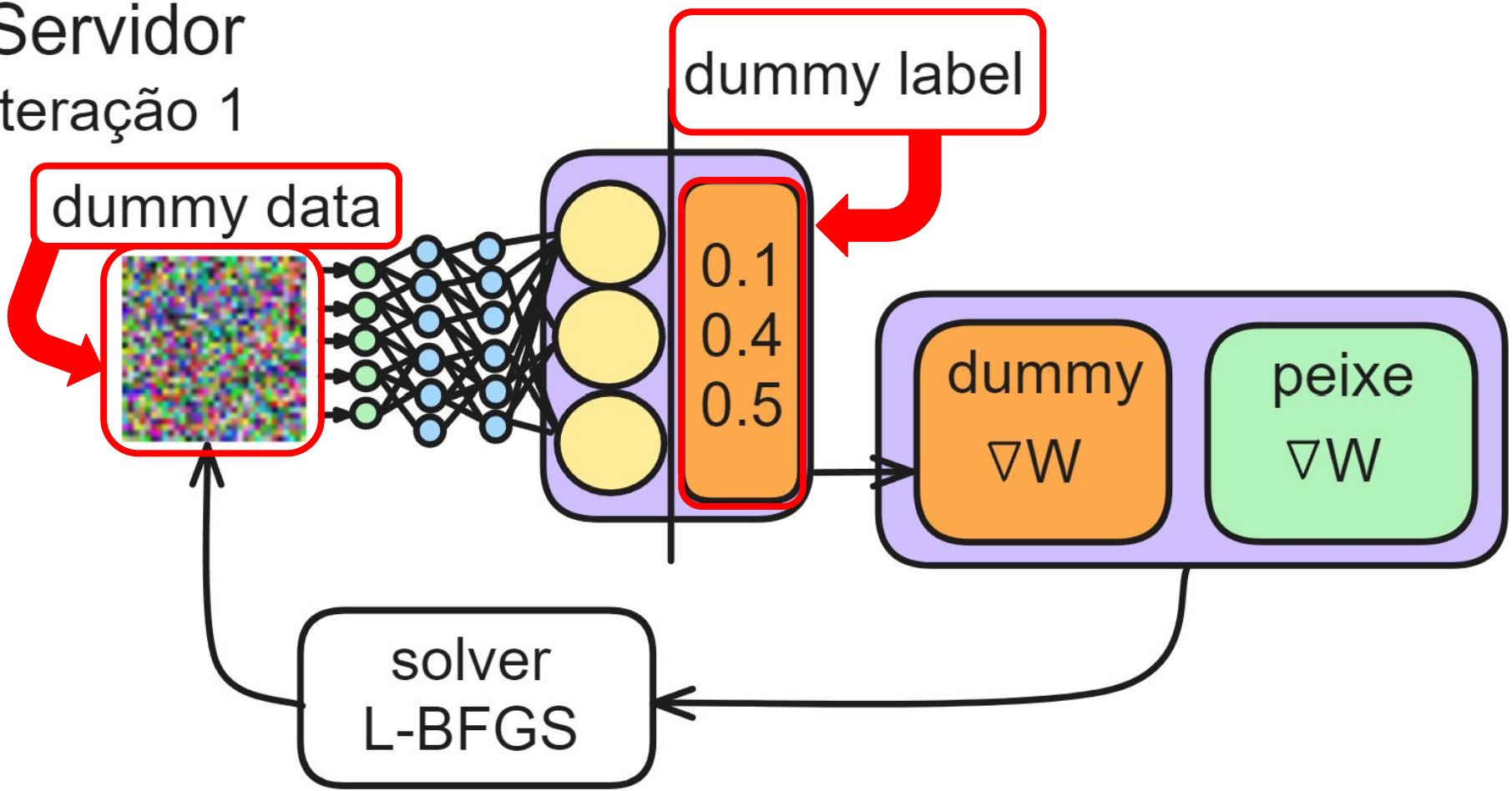
Servidor

Iteração 1



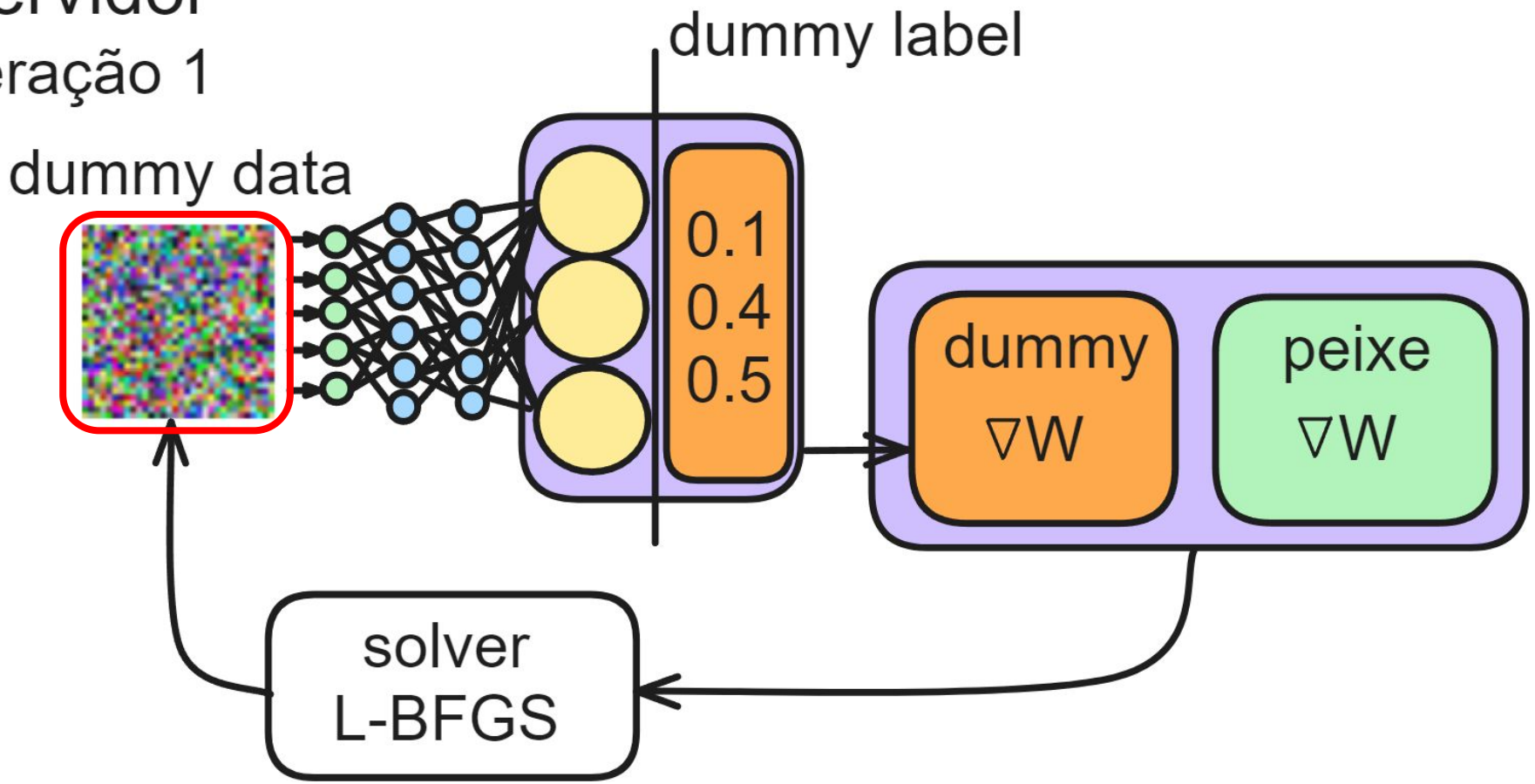
Servidor

Iteração 1



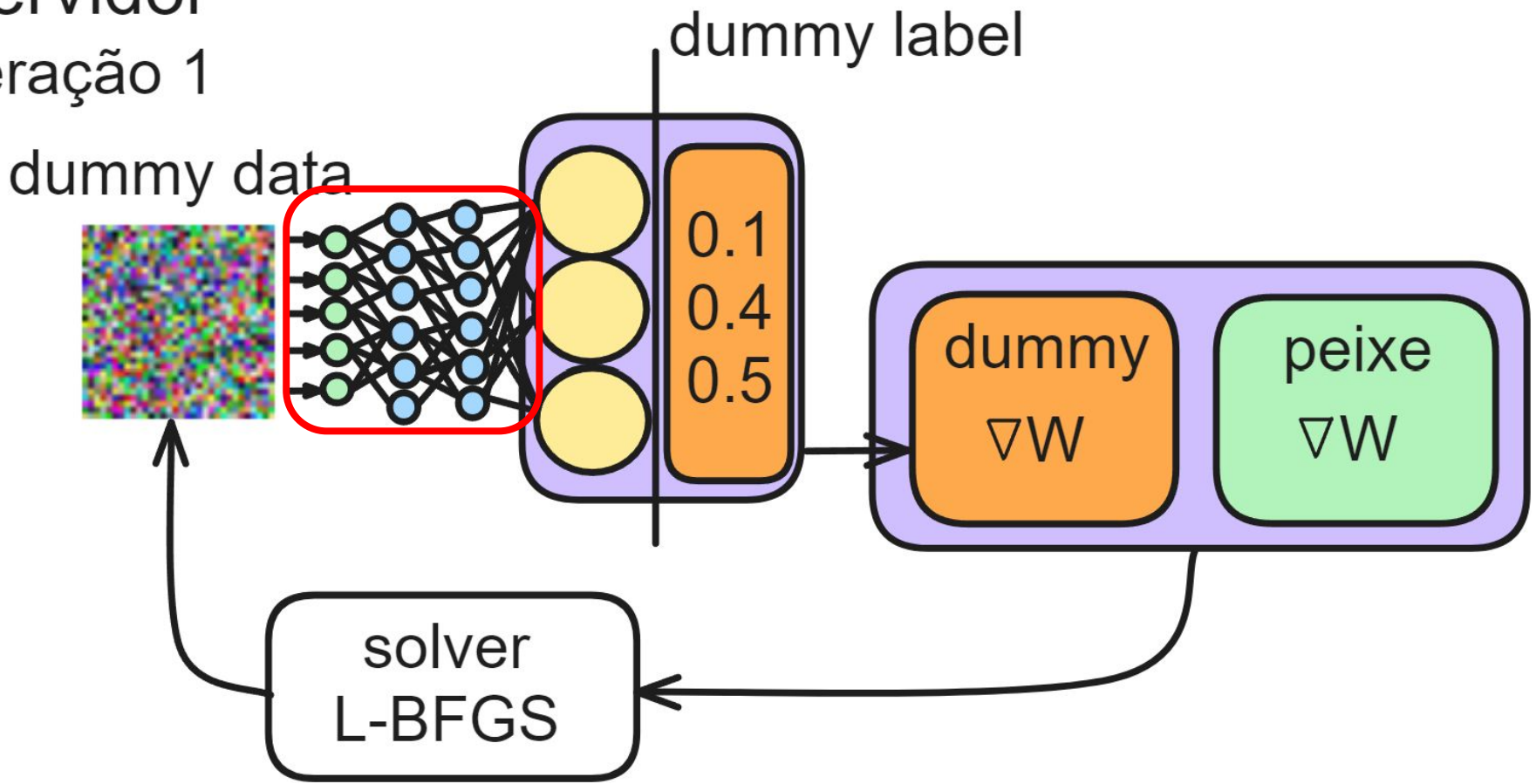
Servidor

Iteração 1



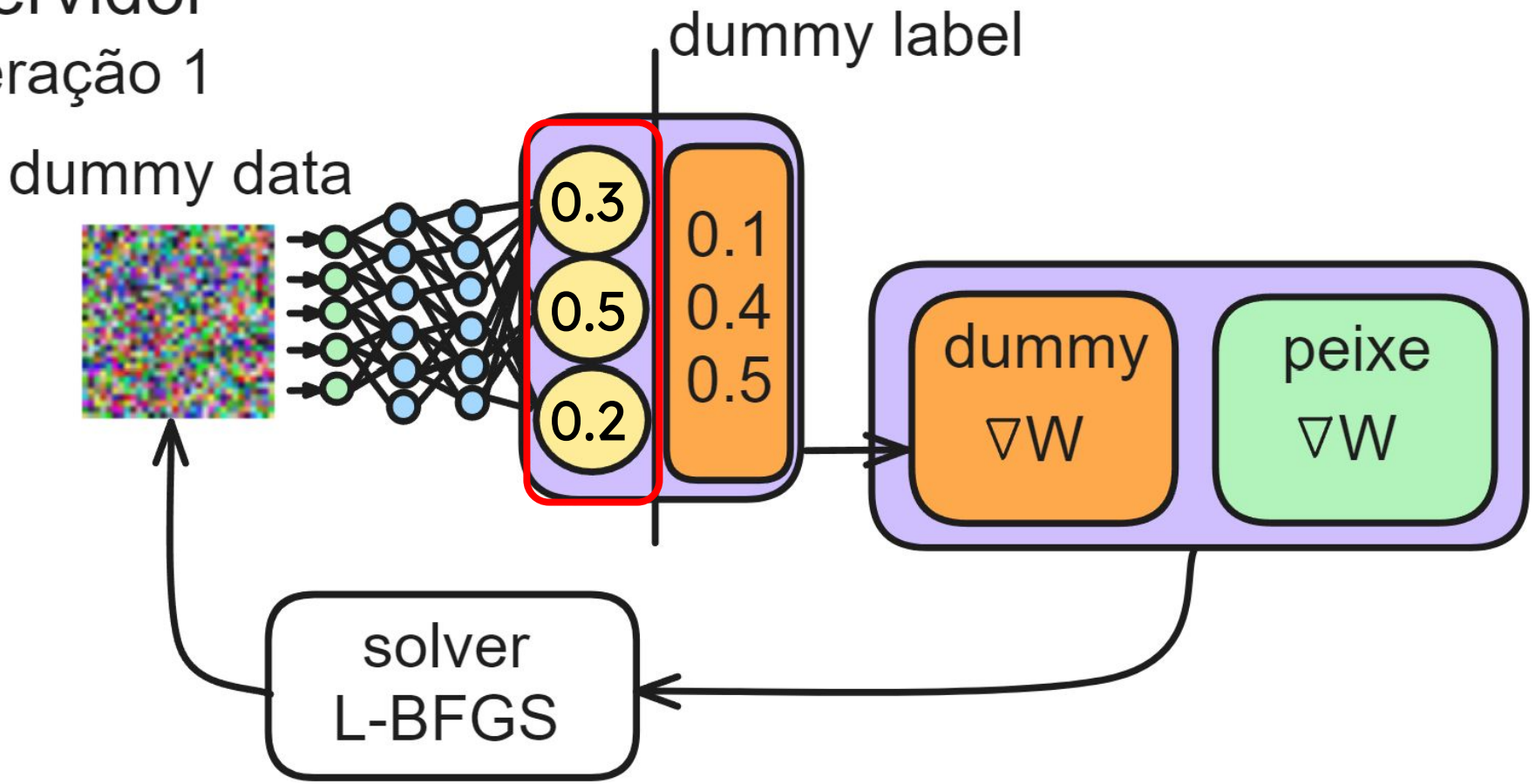
Servidor

Iteração 1



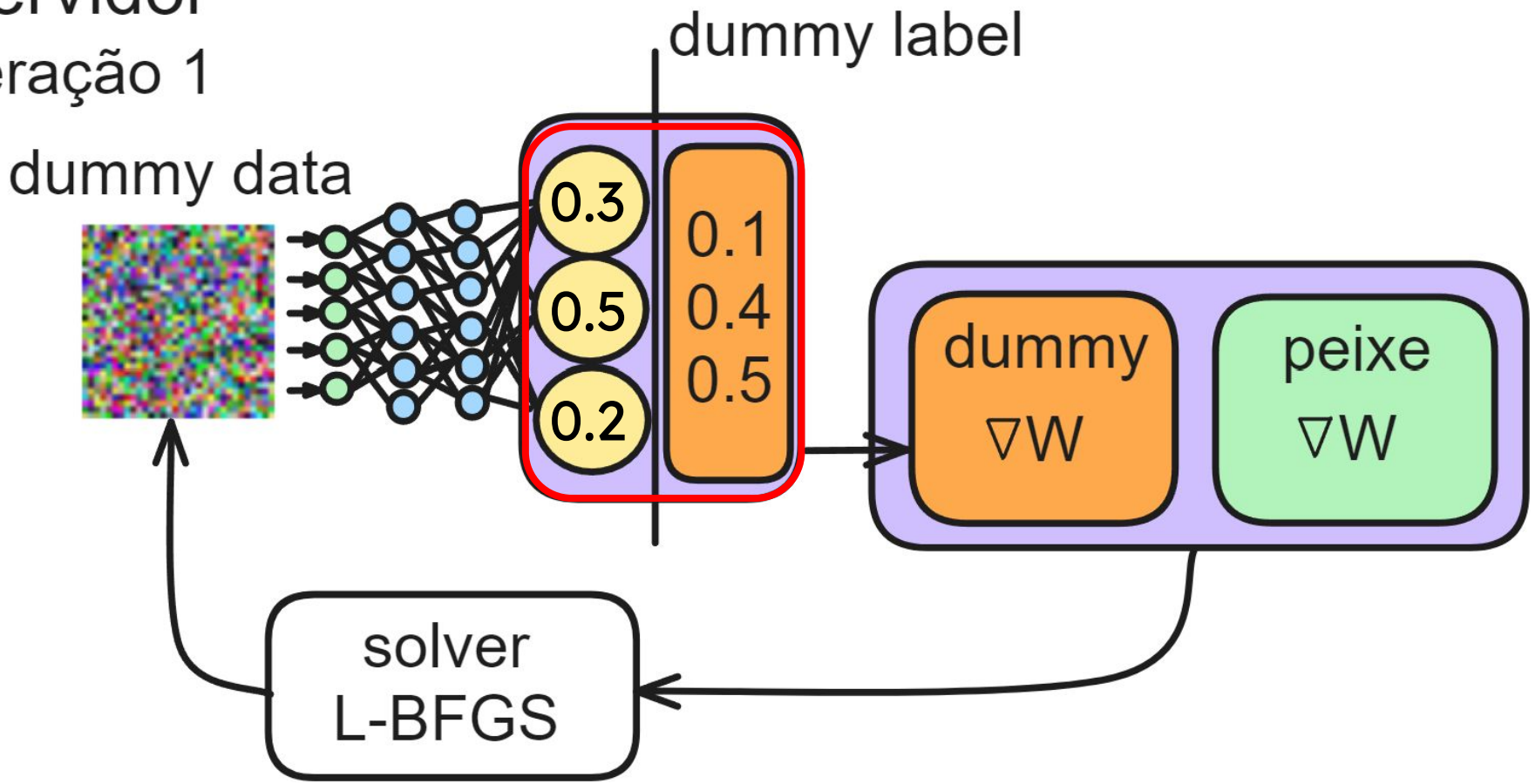
Servidor

Iteração 1



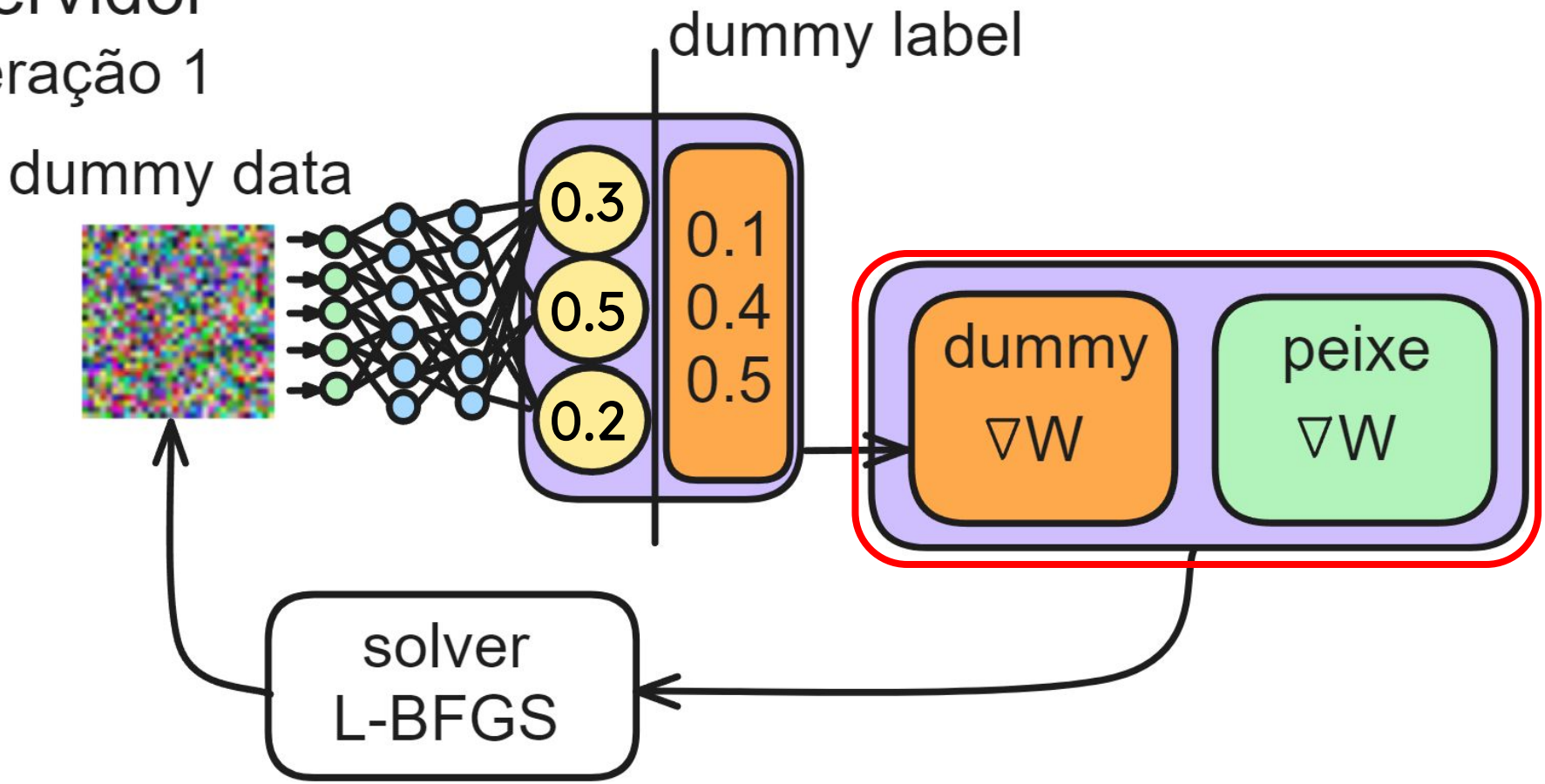
Servidor

Iteração 1



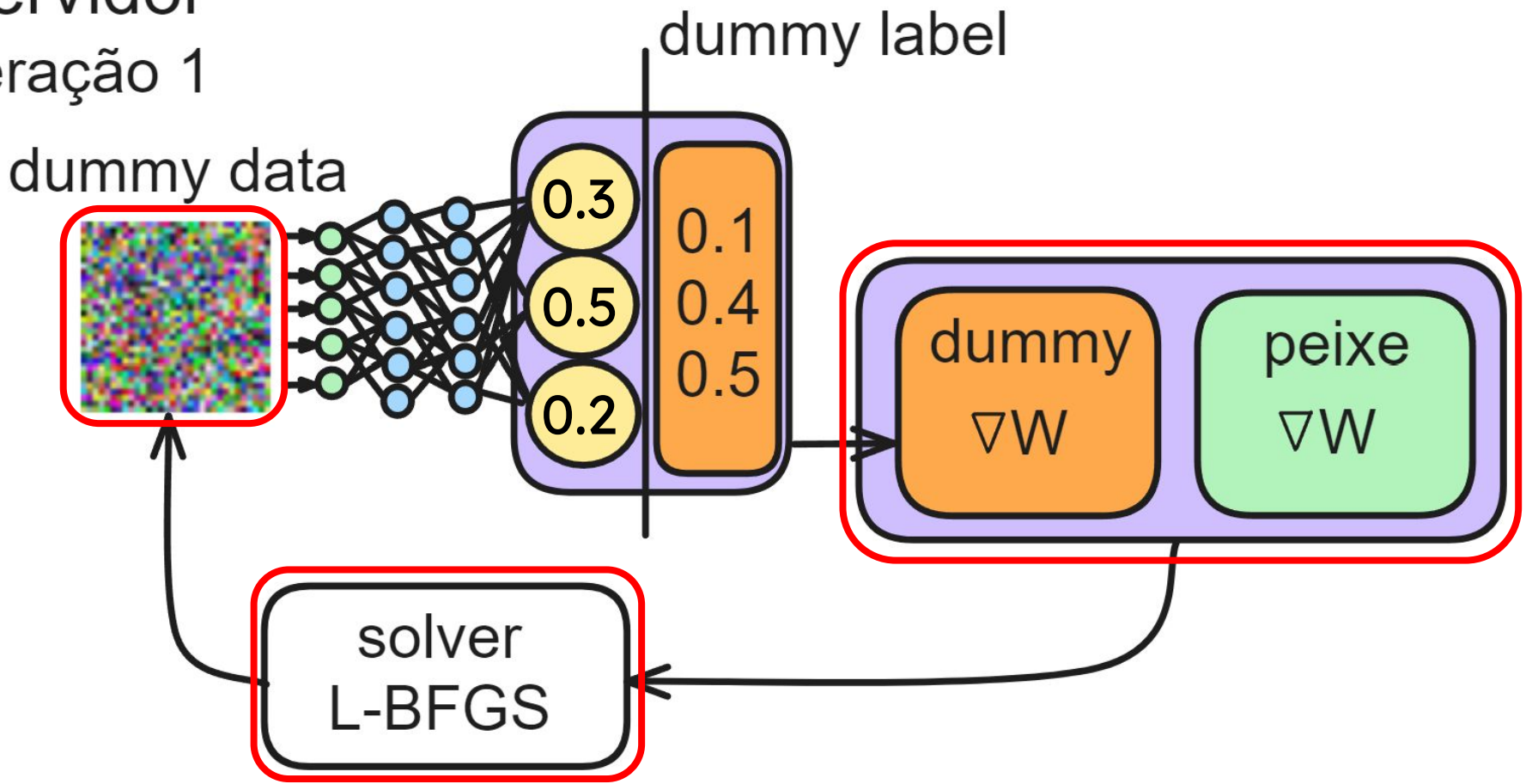
Servidor

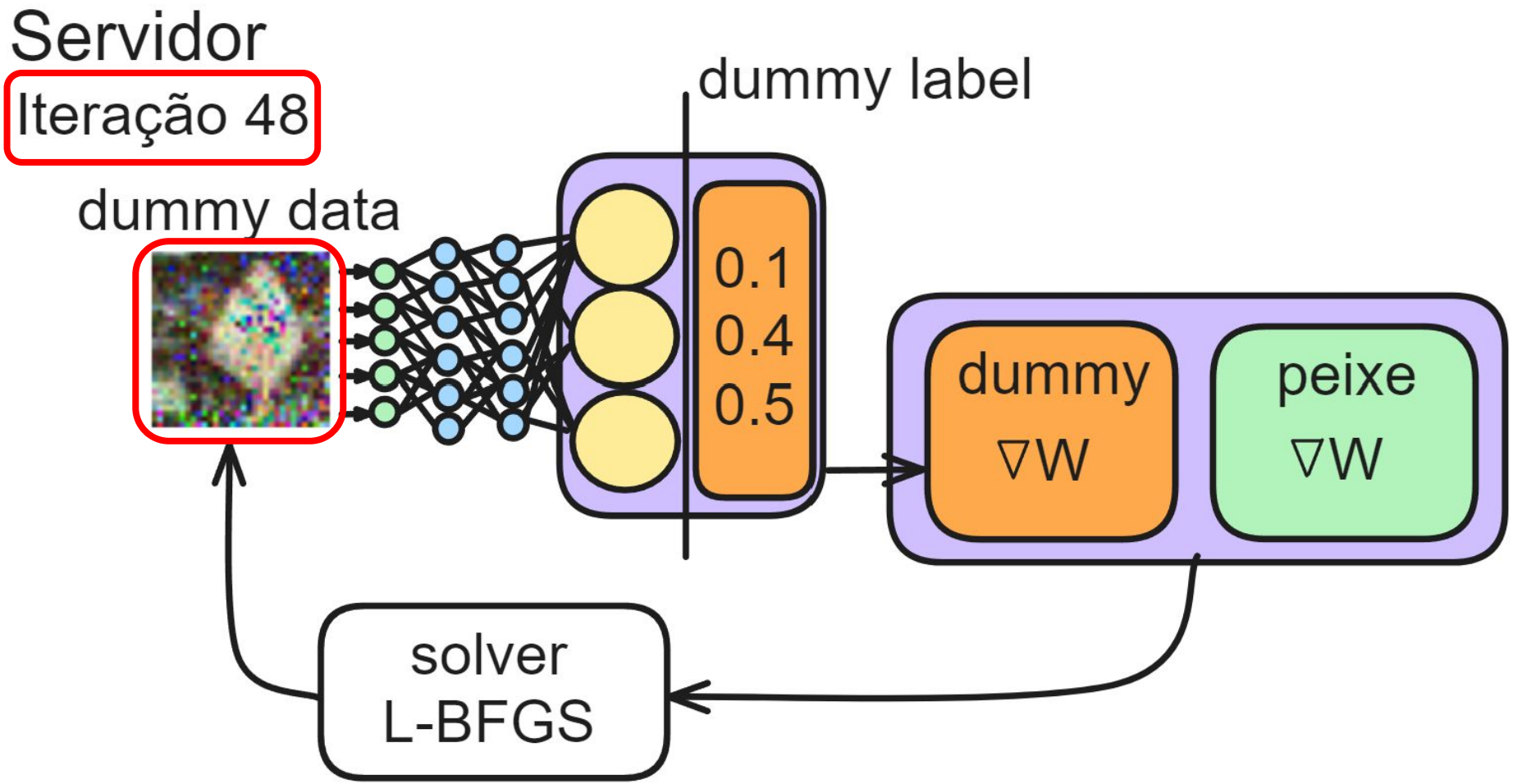
Iteração 1

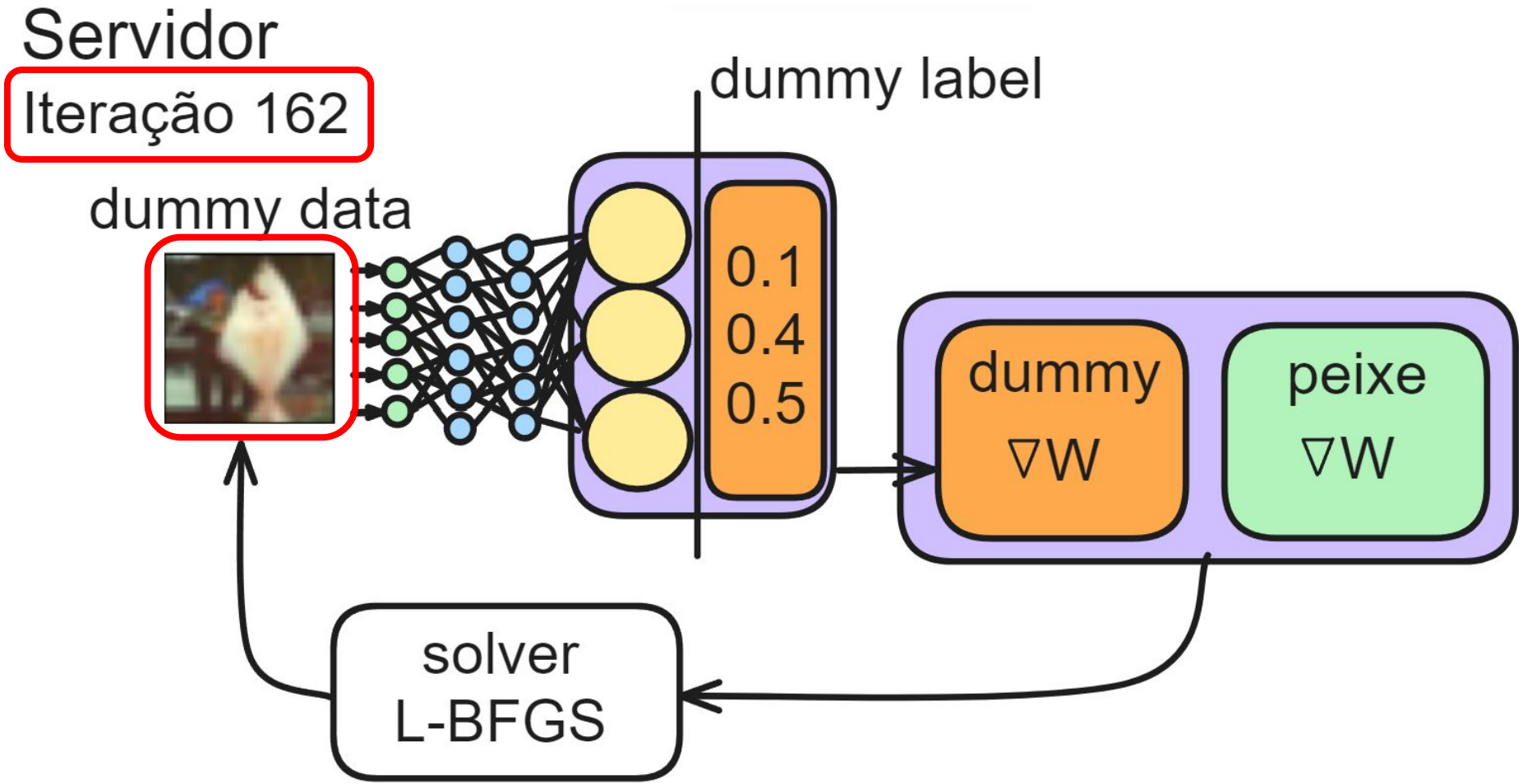


Servidor

Iteração 1

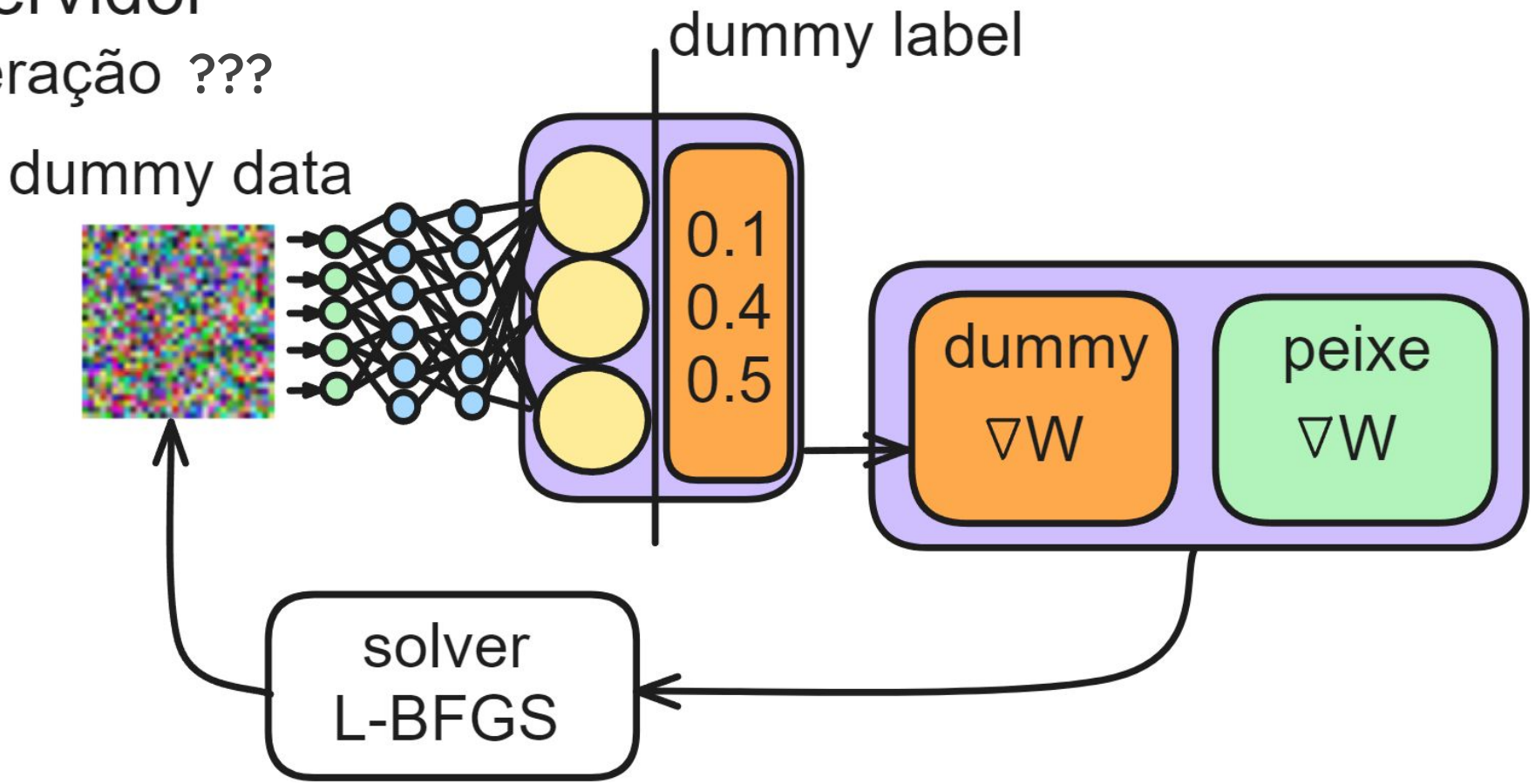



Servidor
Iteração 48
dummy data
dummy label
0.1
0.4
0.5
dummy
∇W
peixe
∇W
solver
L-BFGS


Servidor
Iteração 162
dummy label
dummy data
0.1
0.4
0.5
dummy
∇W
peixe
∇W
solver
L-BFGS

Servidor

Iteração ???



Servidor

Iteração ???

dummy data

dummy label

0.0
0.0
1.0

iDLG

dummy
∇W

peixe
∇W

solver
L-BFGS

Desafio(s)

- Defesas
- Quão perigoso é esse ataque?

Desafio(s)

- Defesas
- Quão perigoso é esse ataque?

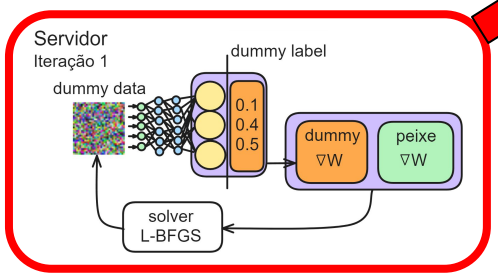
Nova técnica descoberta!

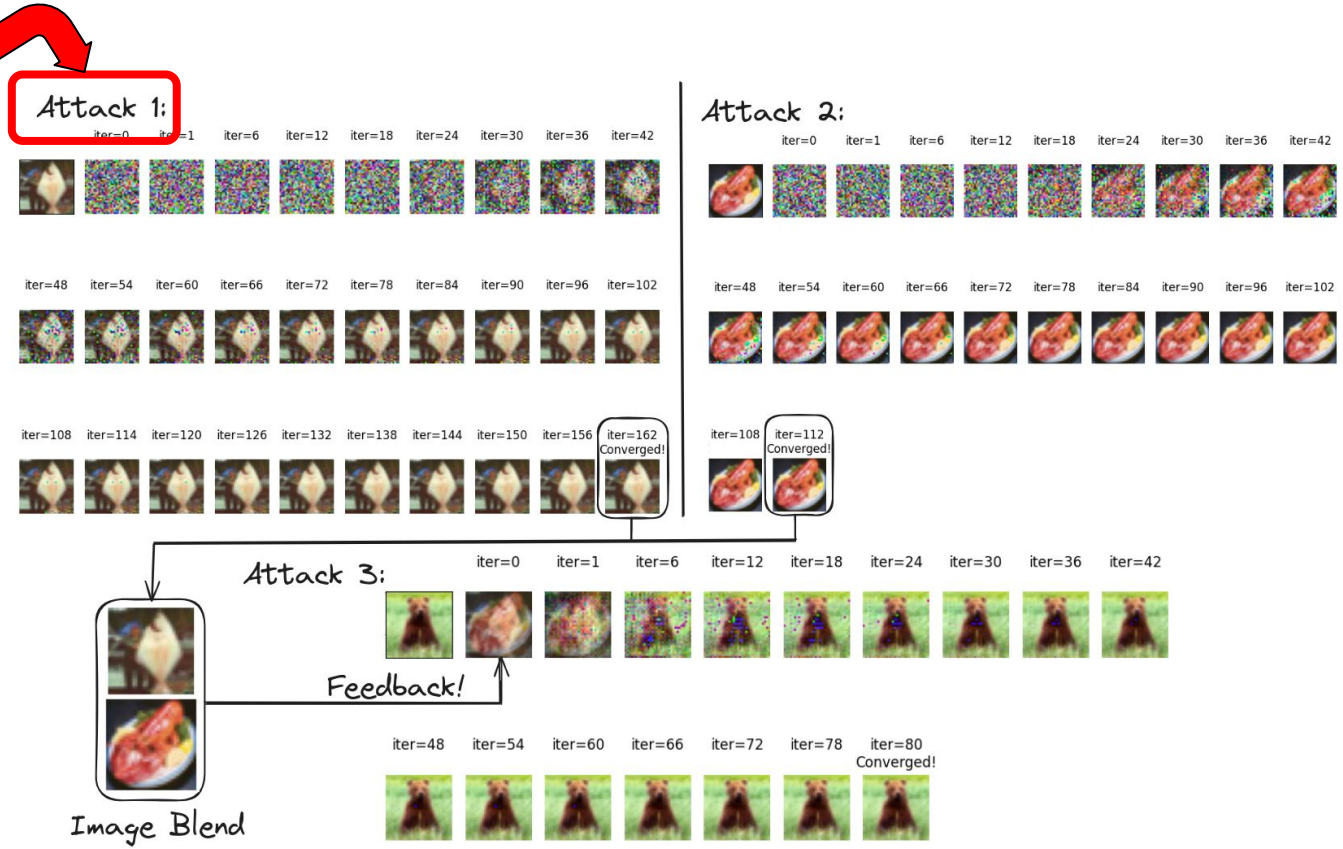
DLG-FB

DLG-FB


Attack 1:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=114 iter=120 iter=126 iter=132 iter=138 iter=144 iter=150 iter=156 iter=162 Converged!
Attack 2:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=112 Converged!
Attack 3:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
Feedback!
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=80 Converged!
Image Blend

DLG-FB


Servidor
Iteração 1
dummy data
dummy label
0.1
0.4
0.5
dummy
∇W
peixe
∇W
solver
L-BFGS


Attack 1:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=114 iter=120 iter=126 iter=132 iter=138 iter=144 iter=150 iter=156 iter=162 Converged!
Attack 2:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=112 Converged!
Attack 3:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
Feedback!
Image Blend
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=80 Converged!

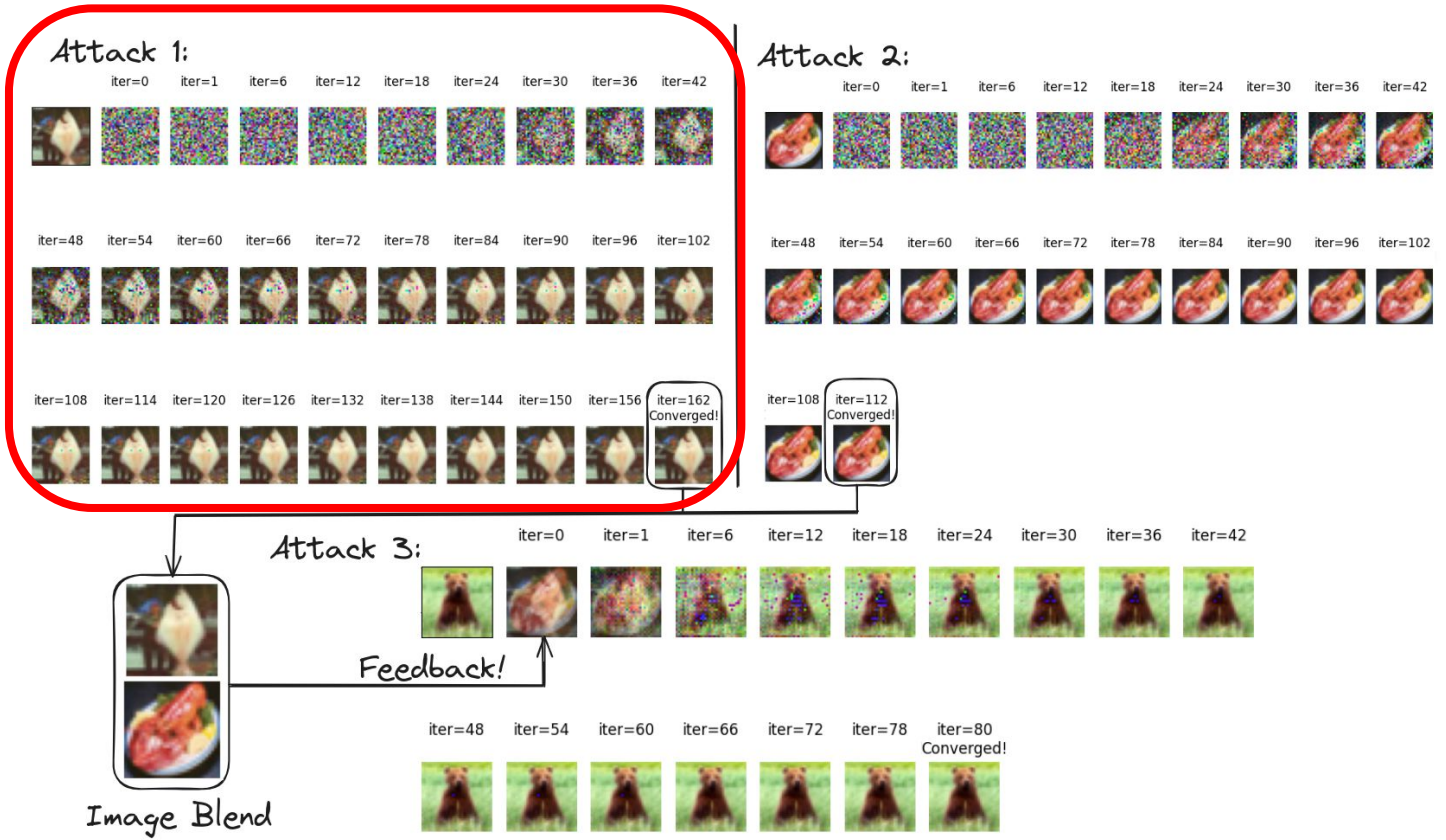
DLG-FB

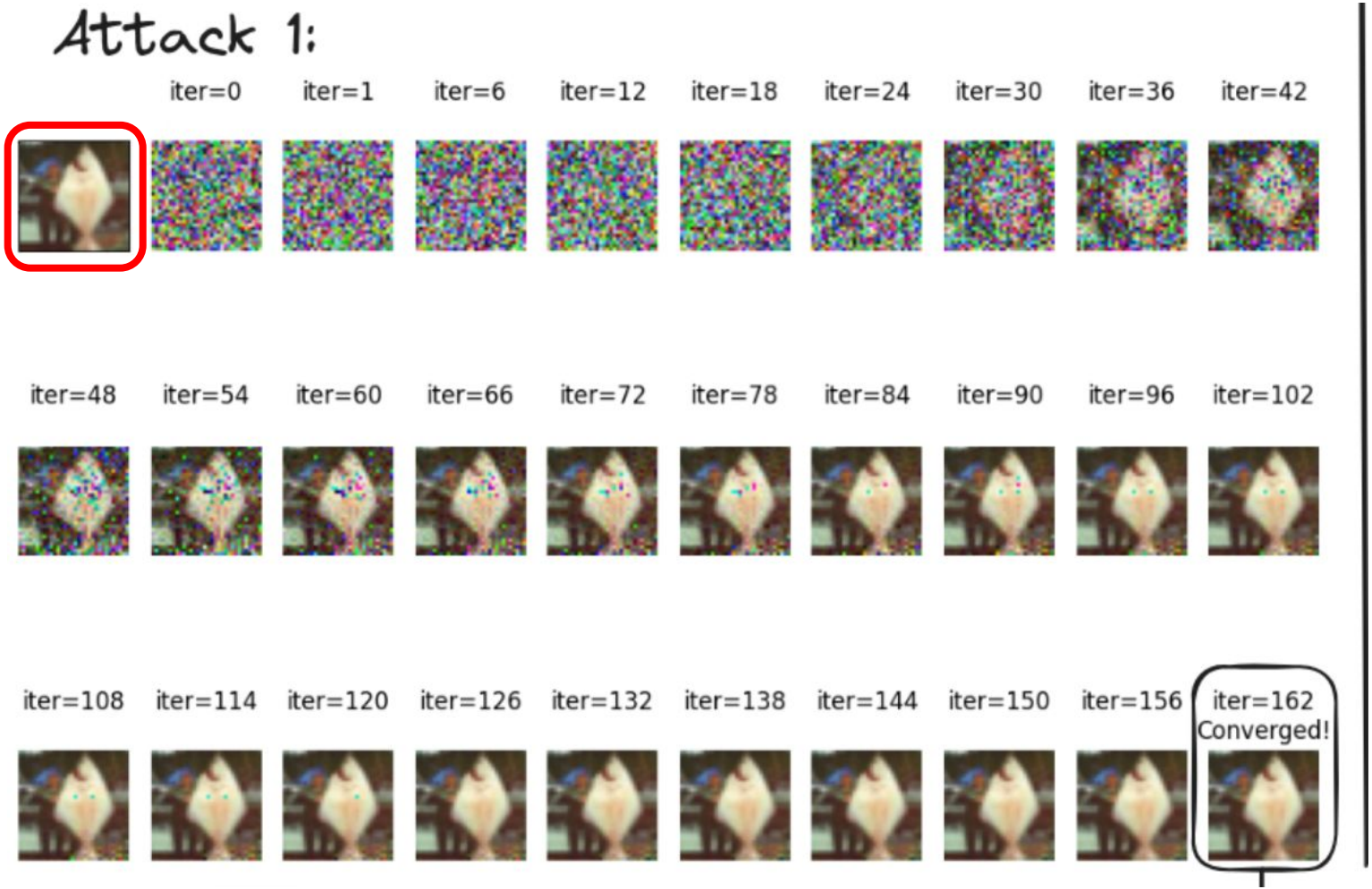

Attack 1:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=114 iter=120 iter=126 iter=132 iter=138 iter=144 iter=150 iter=156 iter=162 Converged!
Attack 2:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=112 Converged!
Attack 3:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
Feedback!
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=80 Converged!
Image Blend

DLG-FB


Attack 1:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=114 iter=120 iter=126 iter=132 iter=138 iter=144 iter=150 iter=156 iter=162 Converged!
Attack 2:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=112 Converged!
Attack 3:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
Feedback!
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=80 Converged!
Image Blend

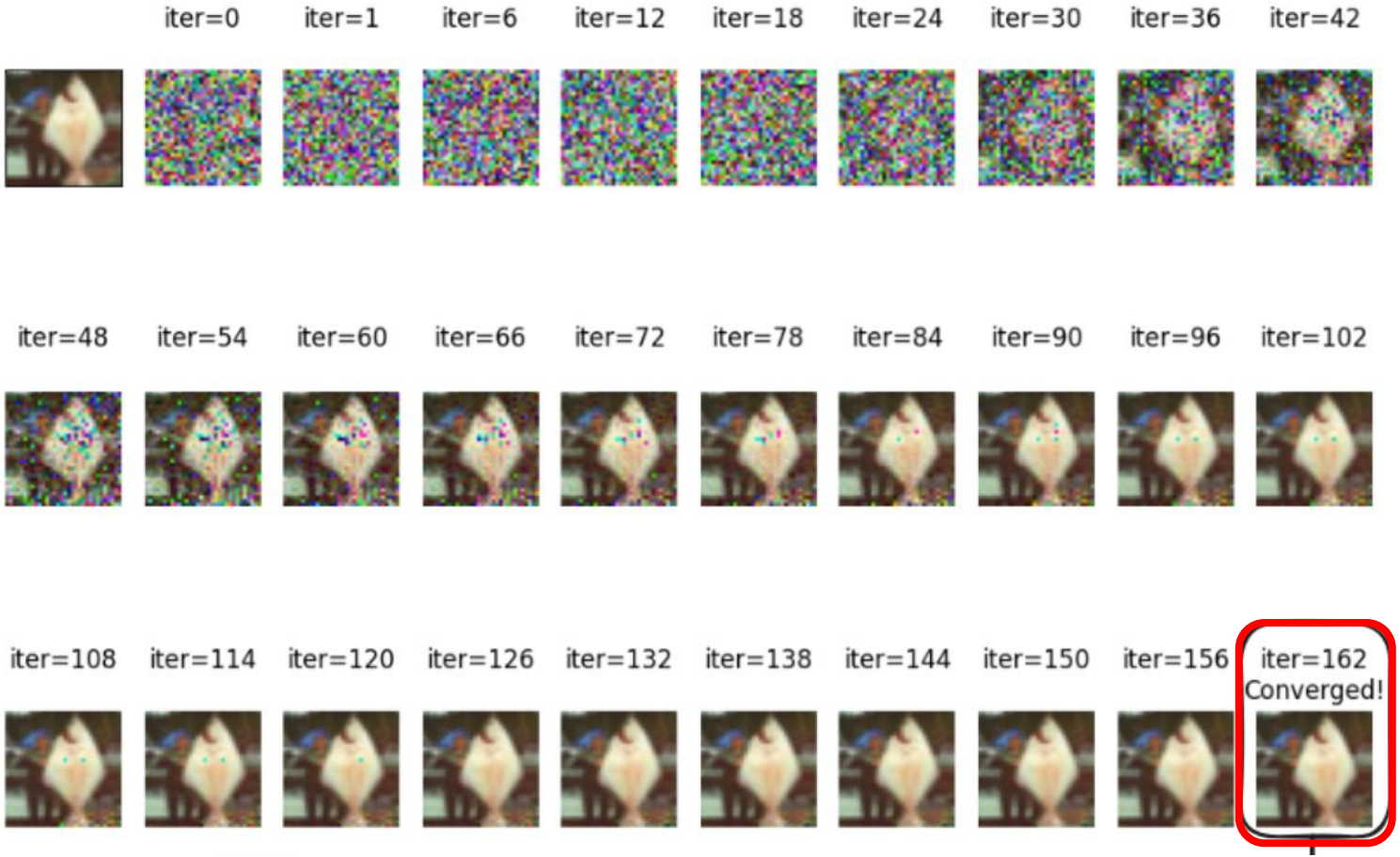
DLG-FB


Attack 1:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=114 iter=120 iter=126 iter=132 iter=138 iter=144 iter=150 iter=156 iter=162 Converged!
Attack 2:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=84 iter=90 iter=96 iter=102
iter=108 iter=112 Converged!
Attack 3:
iter=0 iter=1 iter=6 iter=12 iter=18 iter=24 iter=30 iter=36 iter=42
Feedback!
iter=48 iter=54 iter=60 iter=66 iter=72 iter=78 iter=80 Converged!
Image Blend

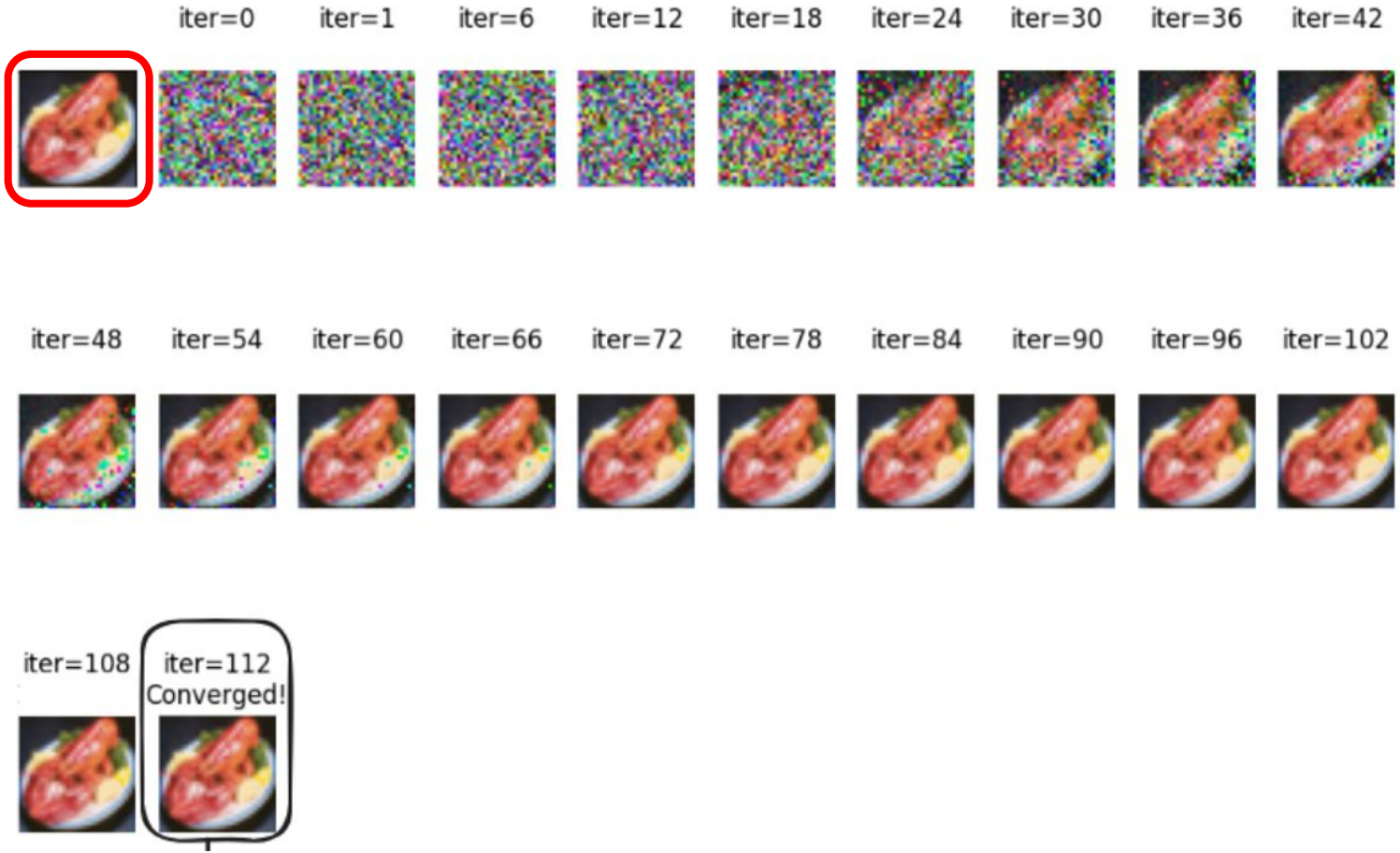

Attack 1:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=84
iter=90
iter=96
iter=102
iter=108
iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!


Attack 1:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=84
iter=90
iter=96
iter=102
iter=108
iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!

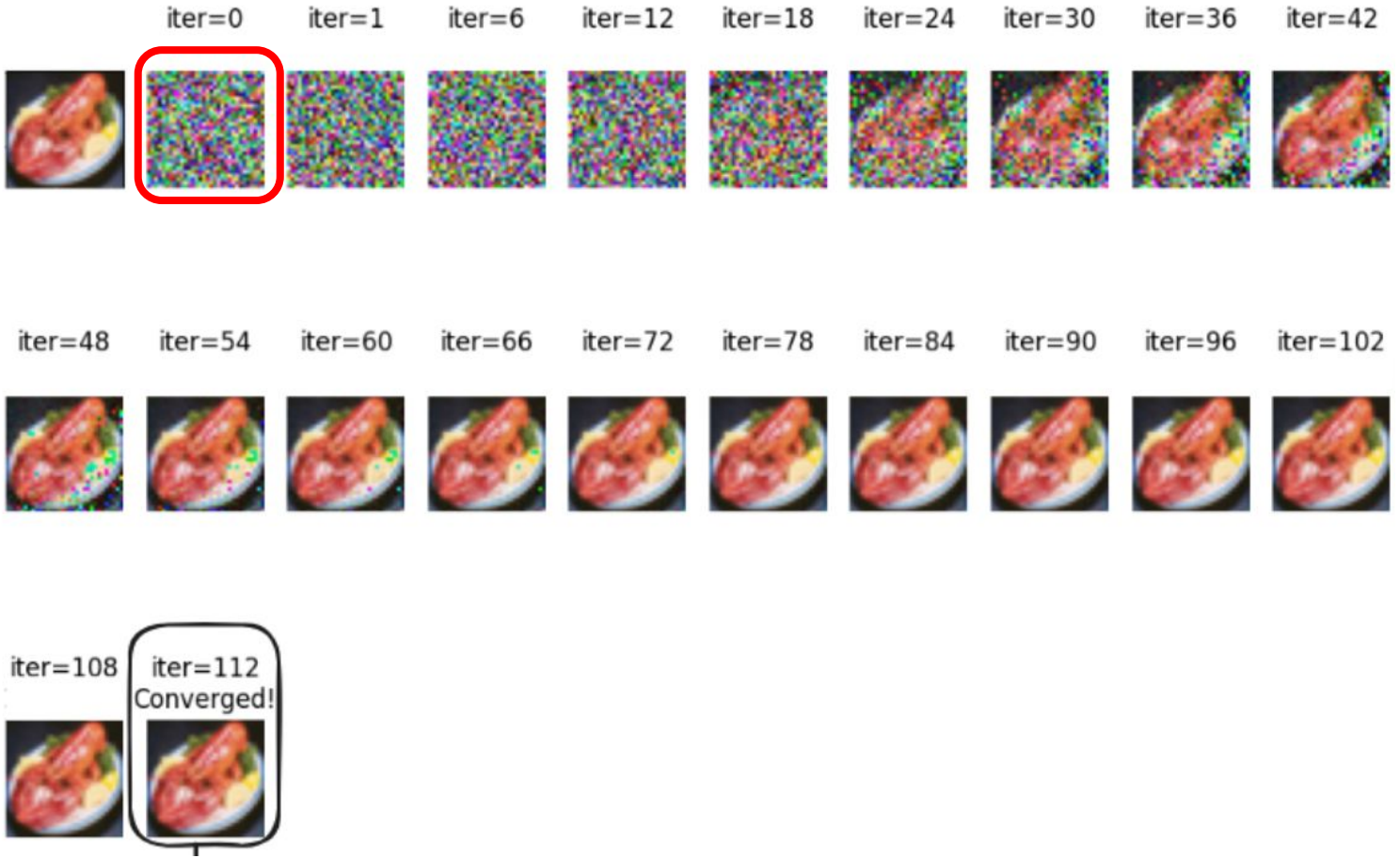
Attack 1:


iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=84
iter=90
iter=96
iter=102
iter=108
iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!

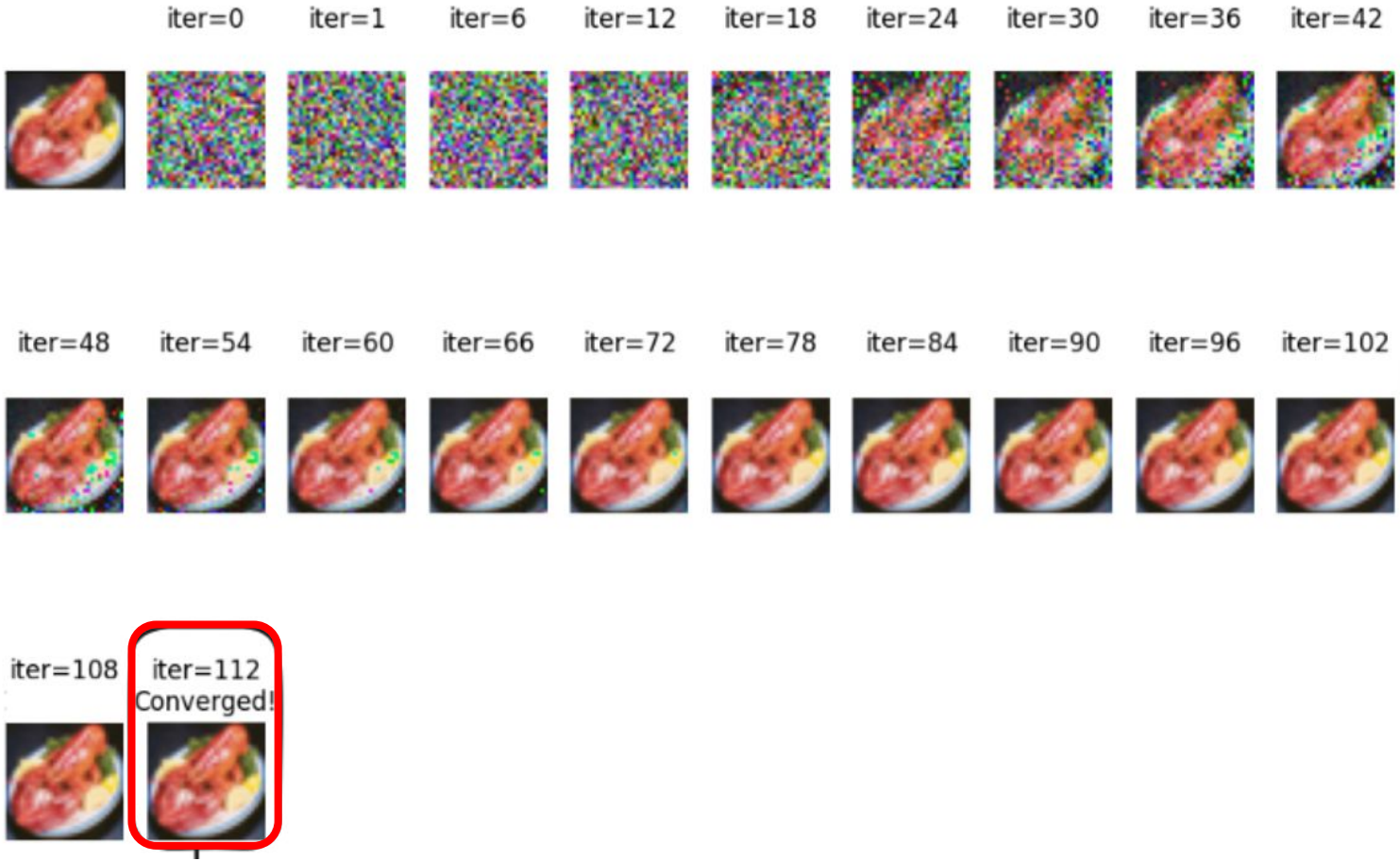
Attack 2:


iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=84
iter=90
iter=96
iter=102
iter=108
iter=112
Converged!

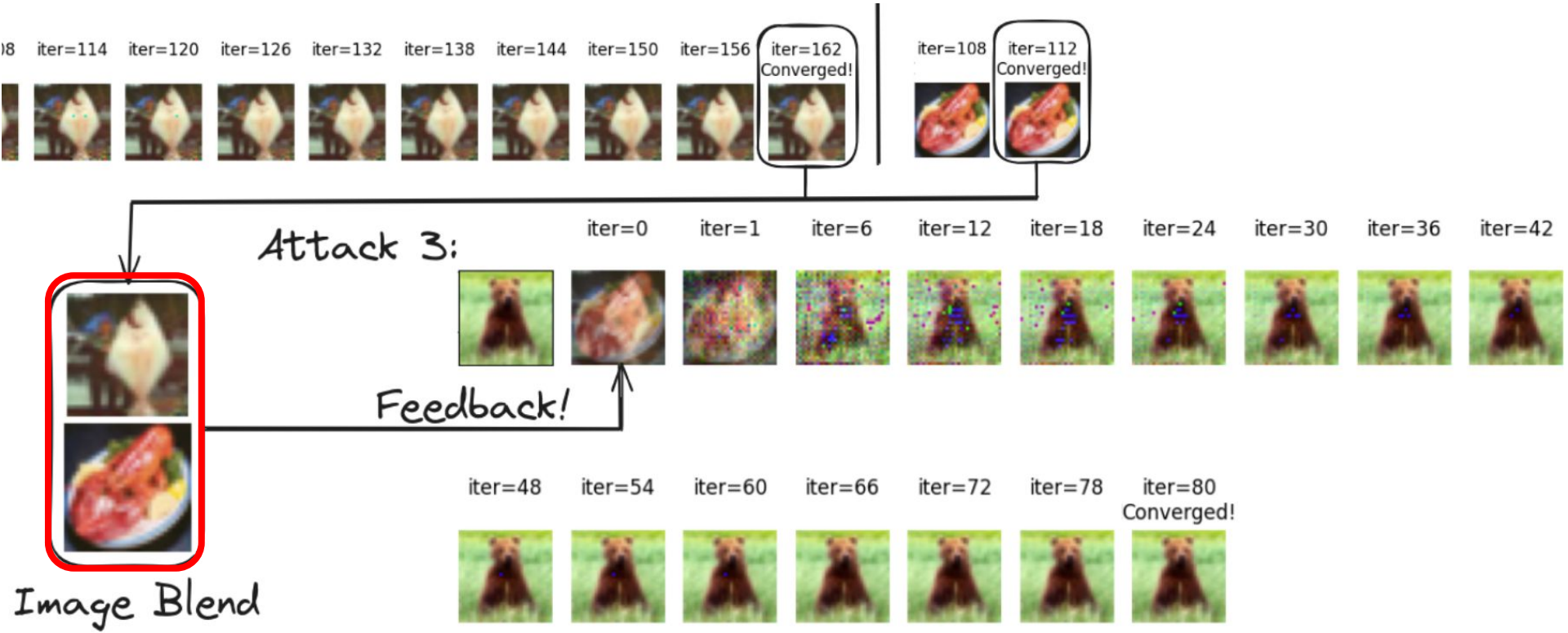
Attack 2:

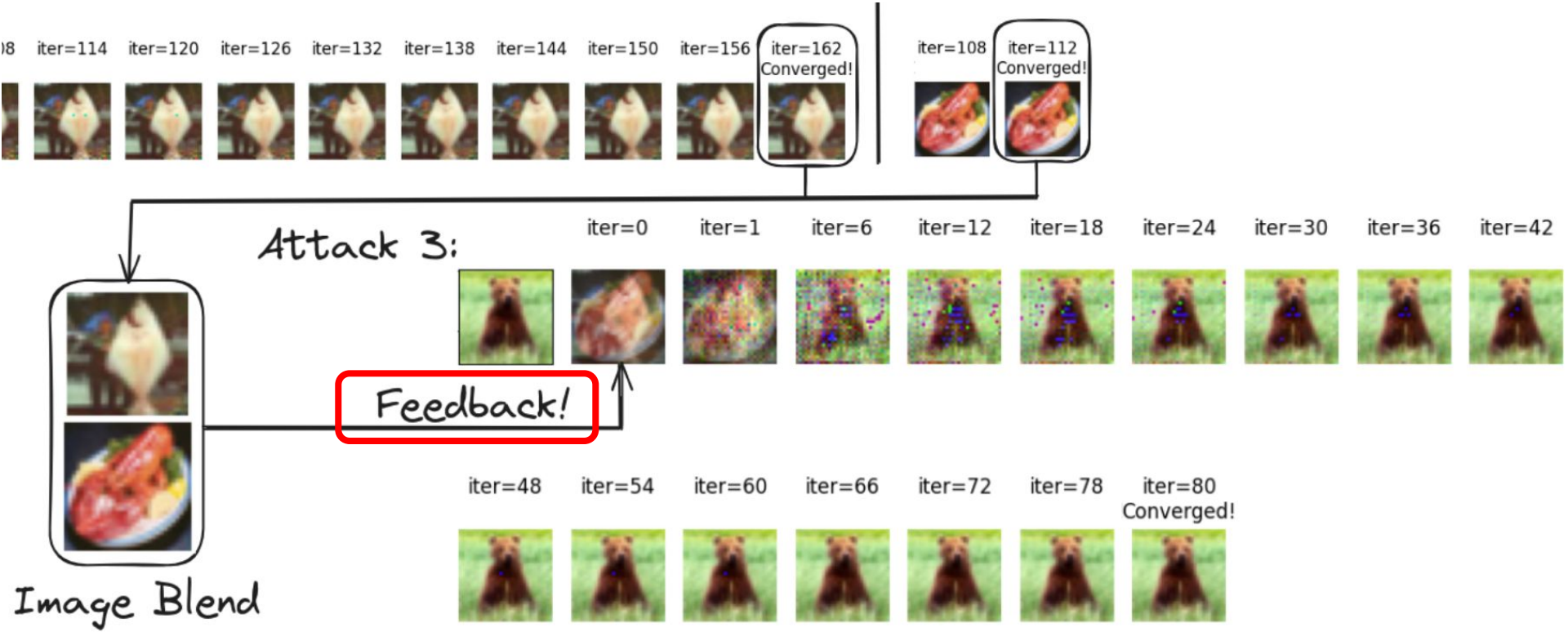

iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=84
iter=90
iter=96
iter=102
iter=108
iter=112
Converged!

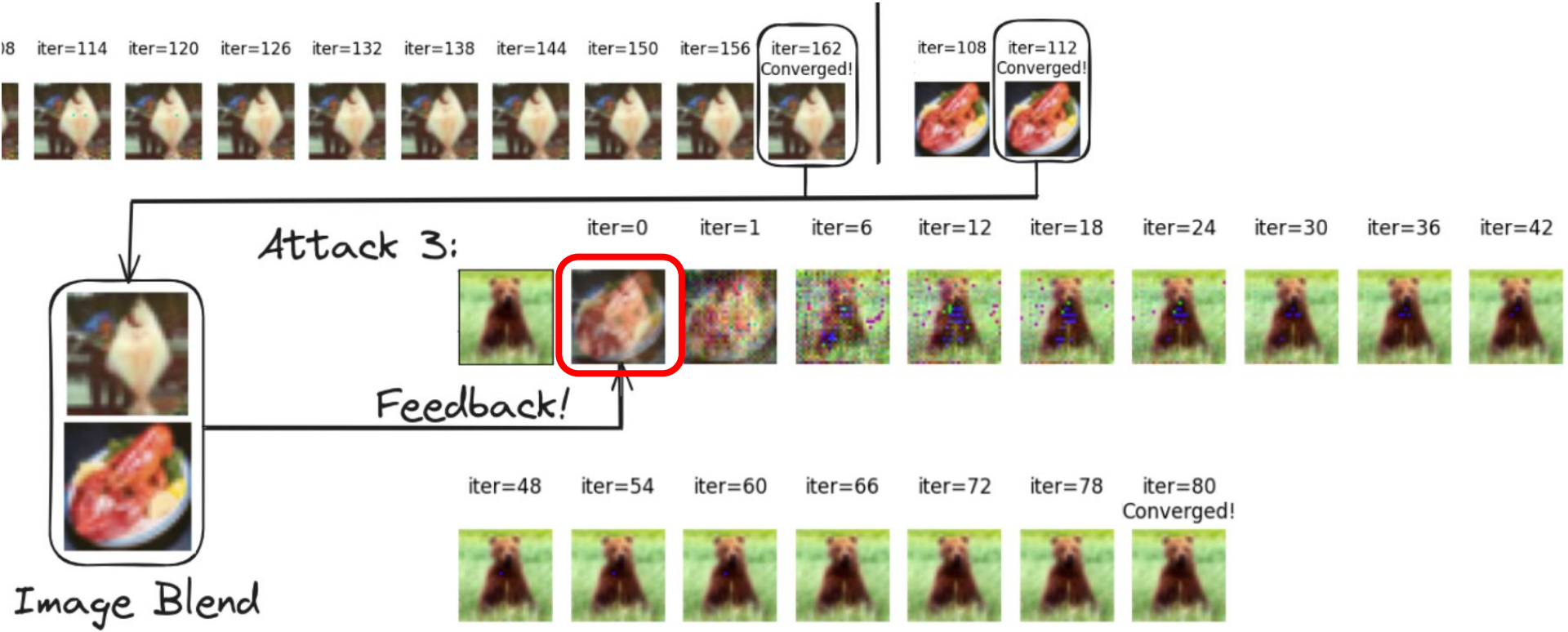
Attack 2:

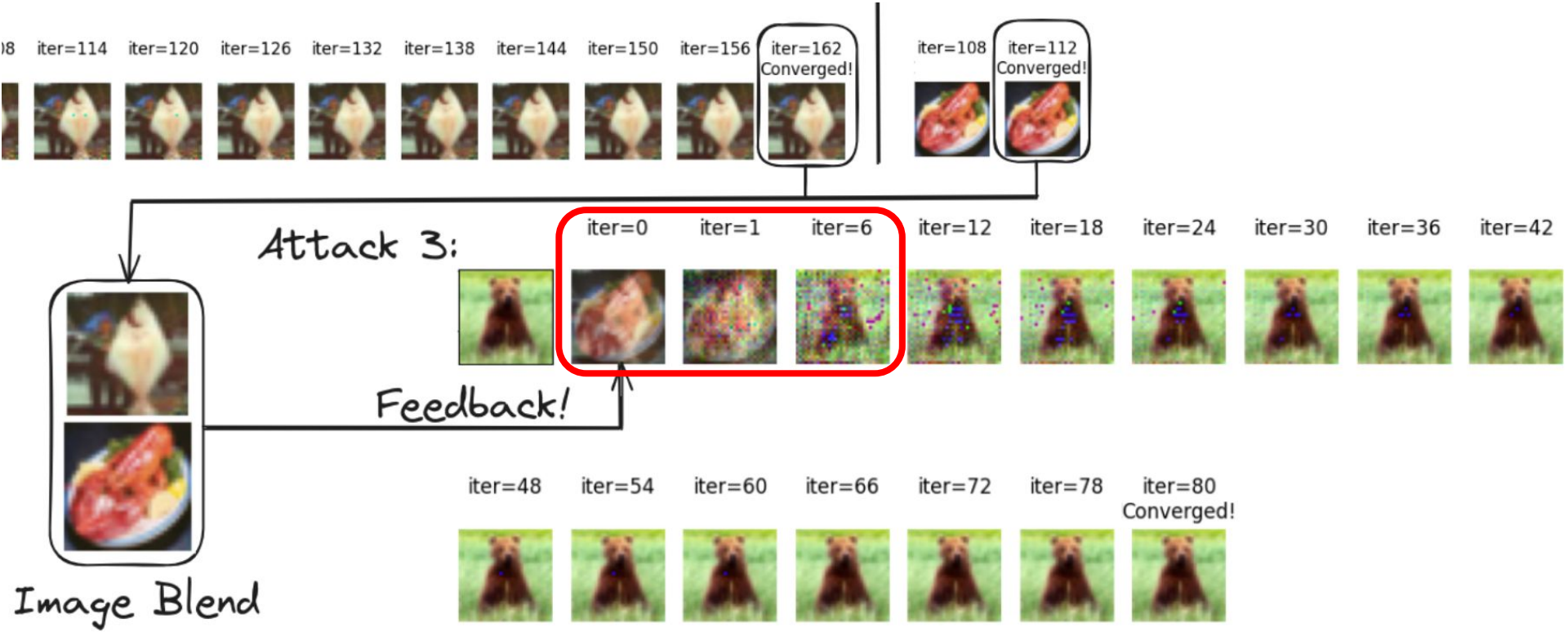

iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=84
iter=90
iter=96
iter=102
iter=108
iter=112
Converged!

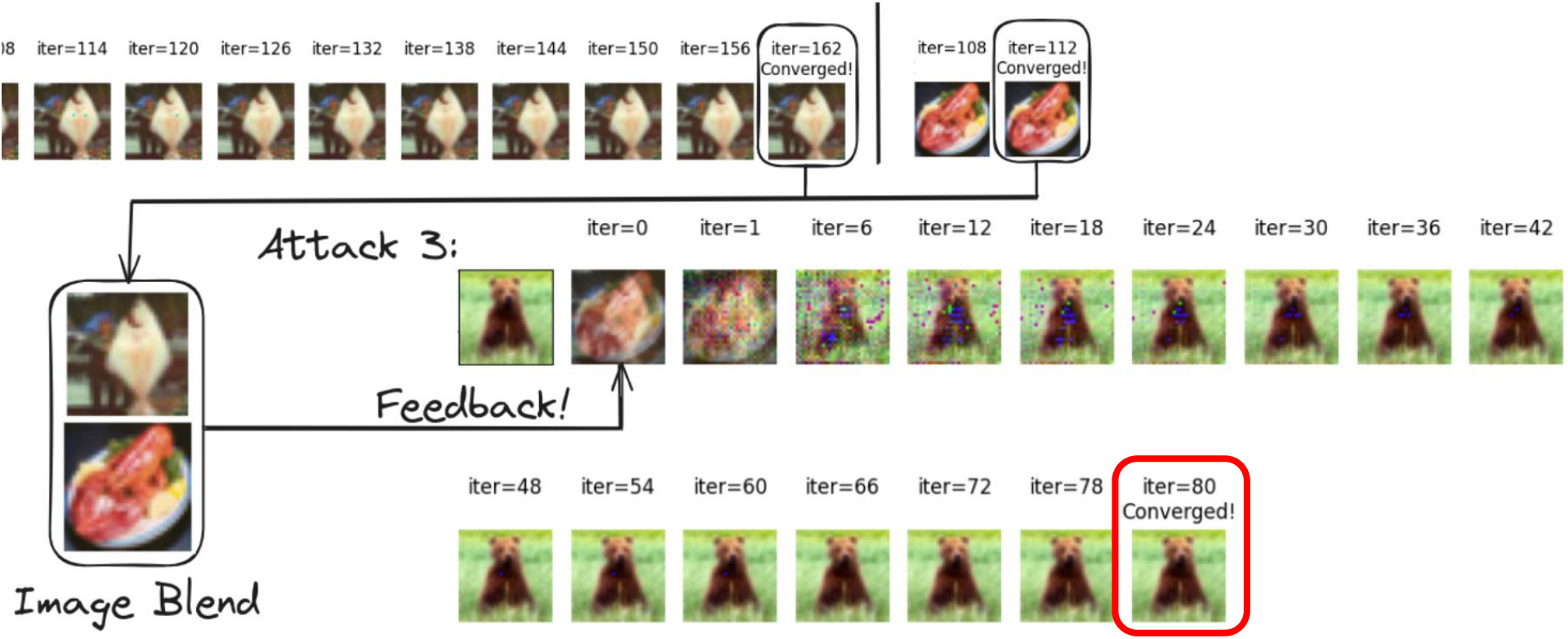

iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!

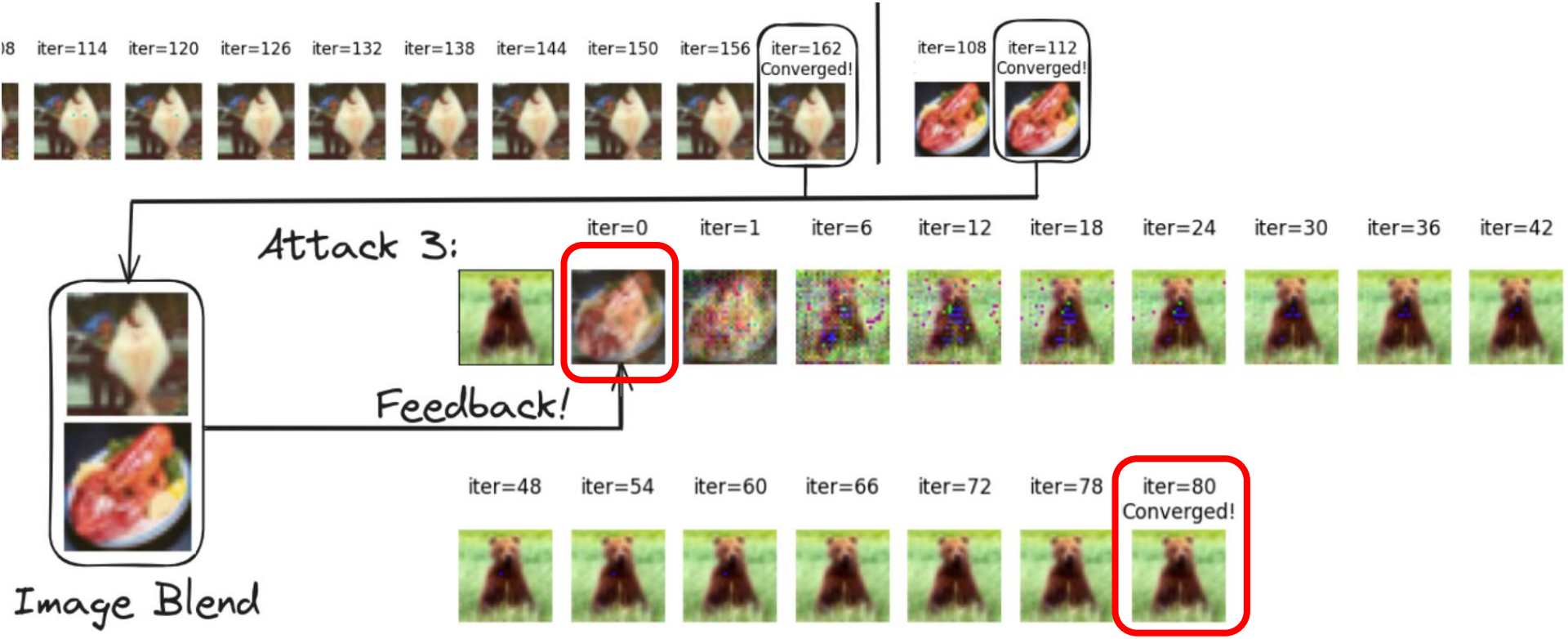

iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!

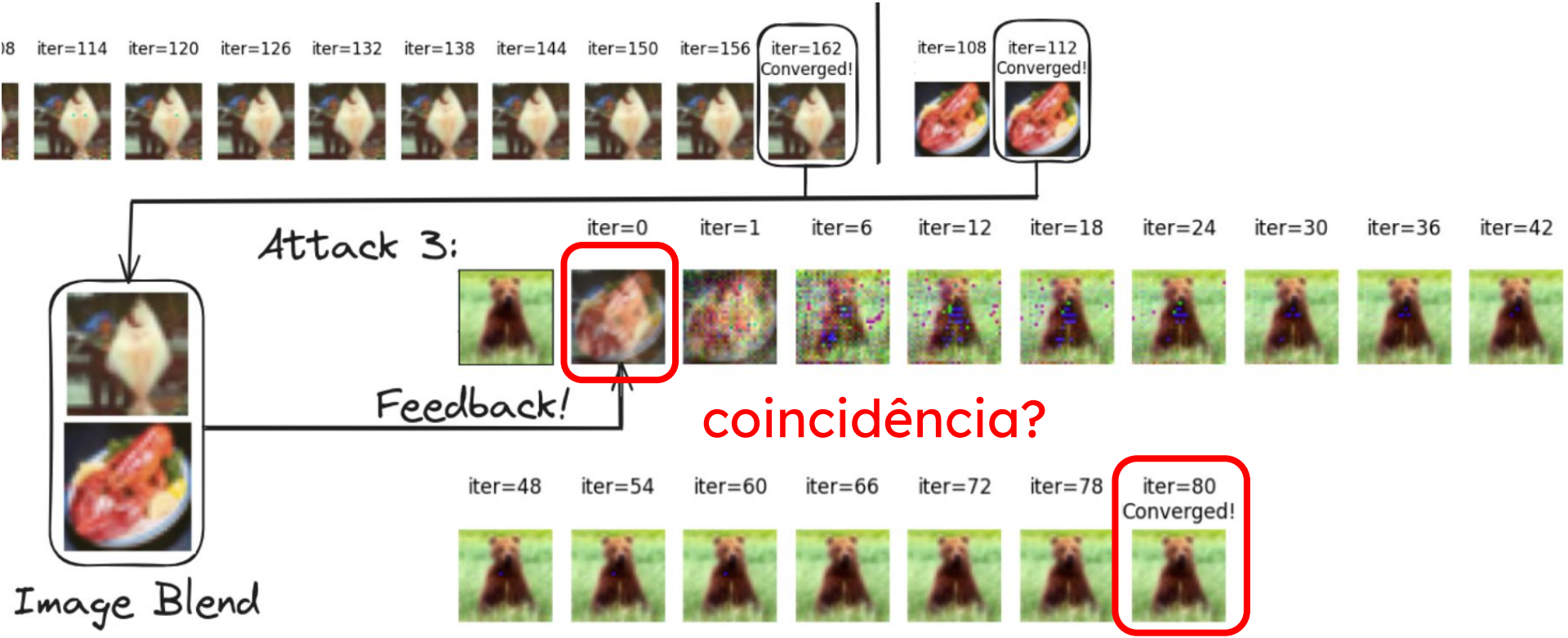

)8
iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!
Image Blend


iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!

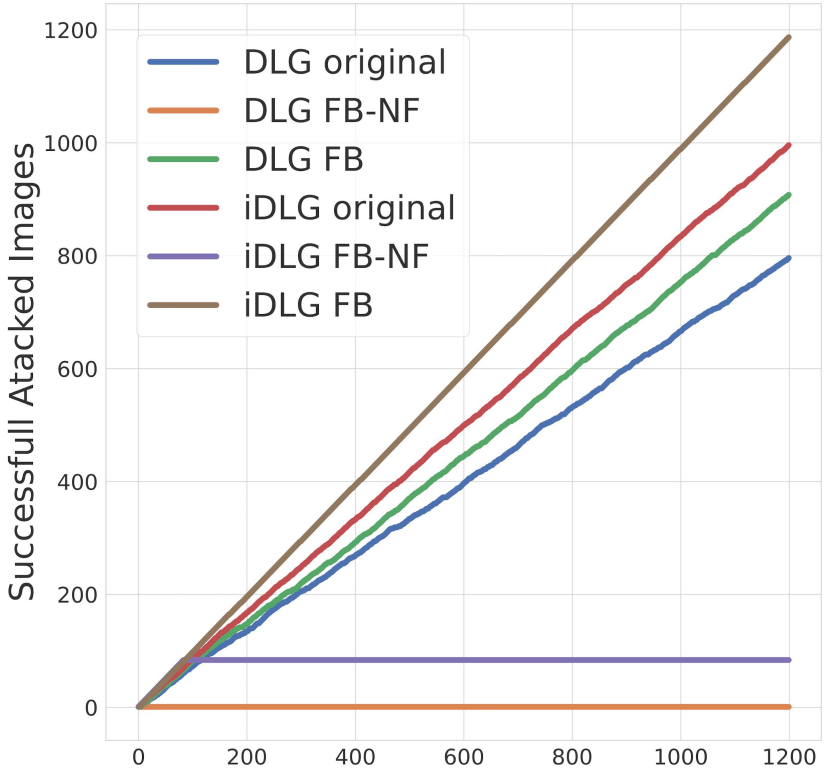

iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!


)8
iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!

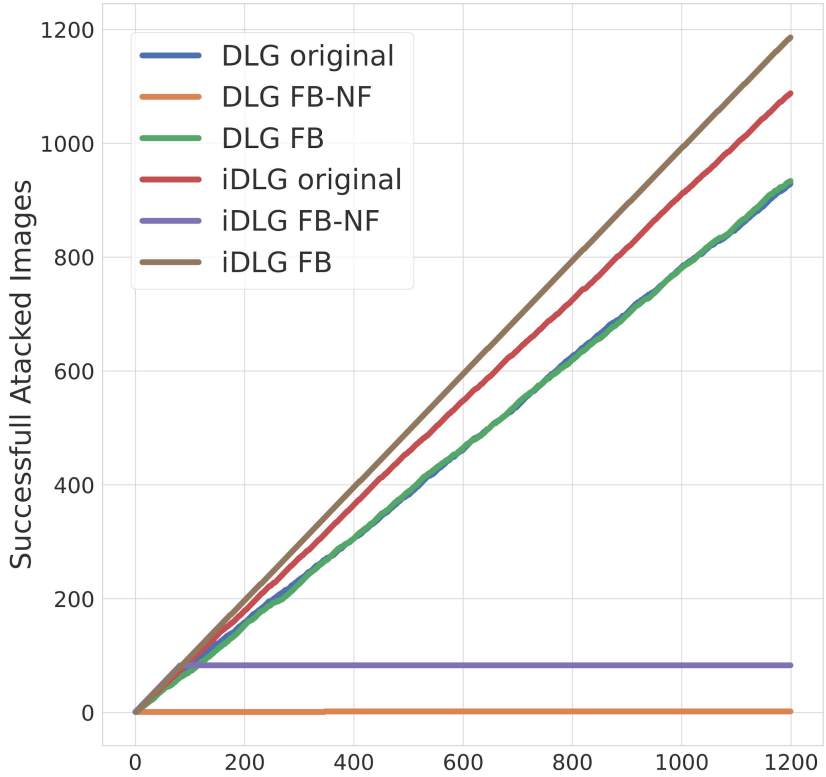

)8
iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!


iter=114
iter=120
iter=126
iter=132
iter=138
iter=144
iter=150
iter=156
iter=162
Converged!
iter=108
iter=112
Converged!
Attack 3:
iter=0
iter=1
iter=6
iter=12
iter=18
iter=24
iter=30
iter=36
iter=42
Feedback!
coincidência?
Image Blend
iter=48
iter=54
iter=60
iter=66
iter=72
iter=78
iter=80
Converged!

Avaliação

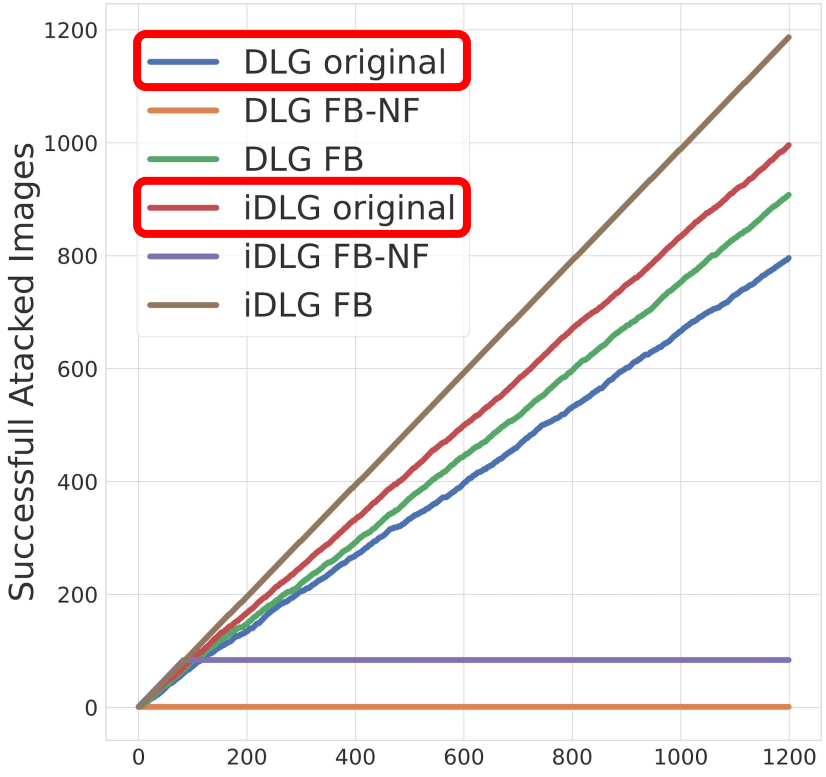


CIFAR100

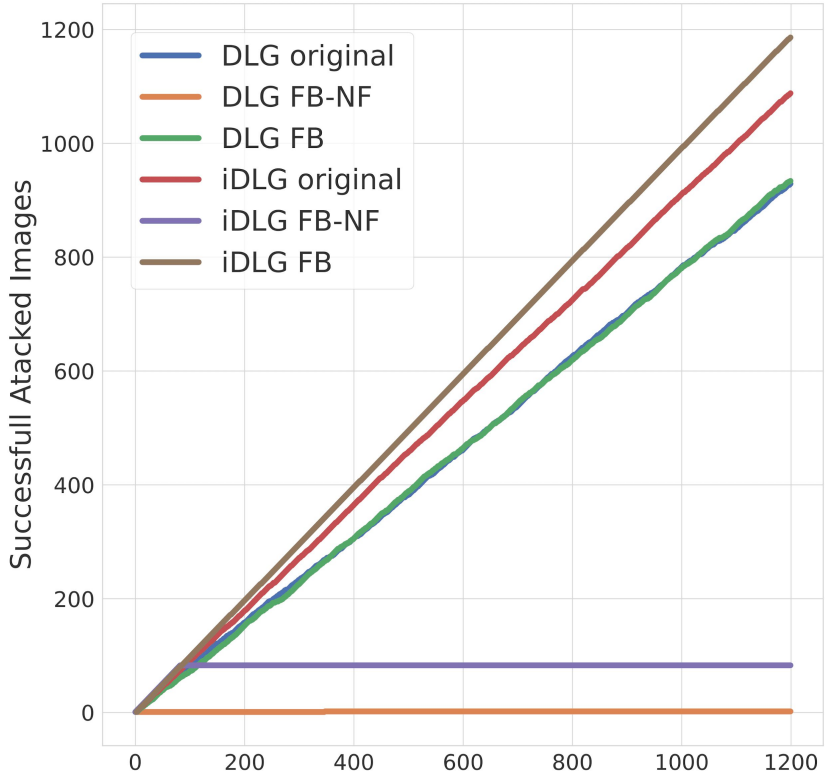


MNIST

Avaliação

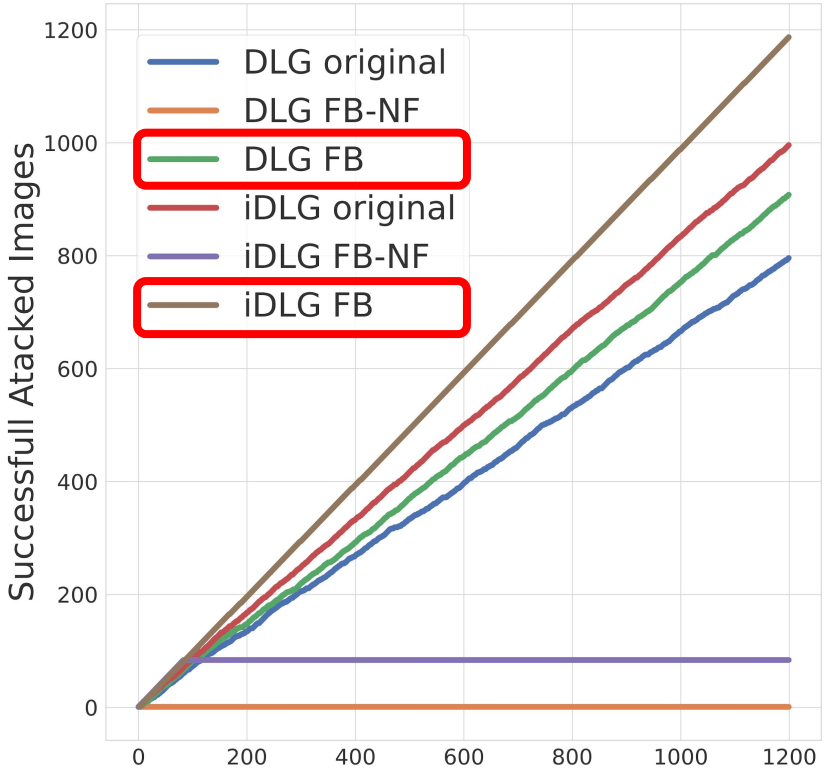


CIFAR100

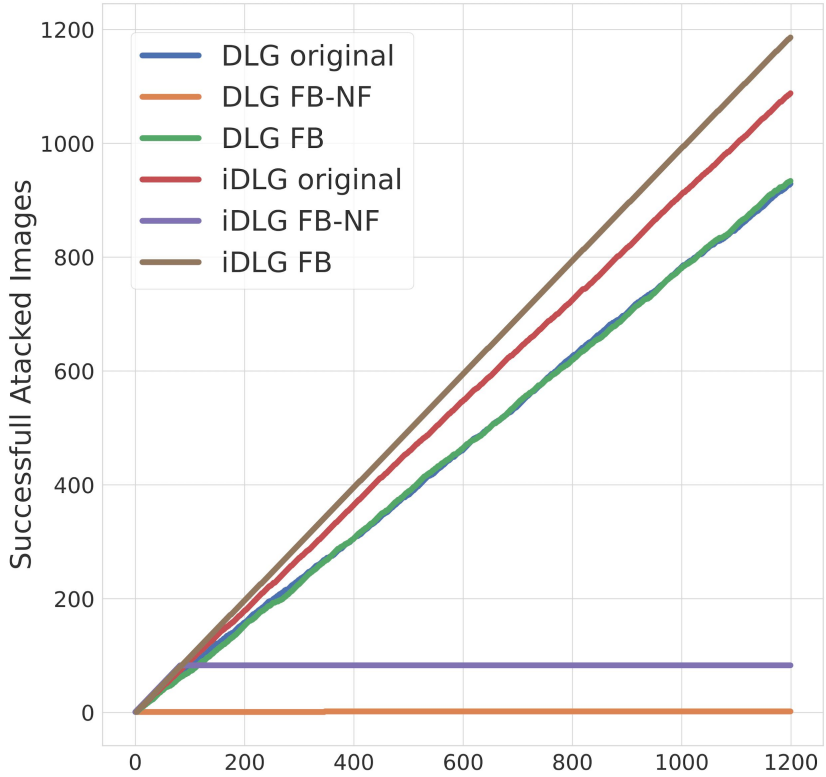


MNIST

Avaliação

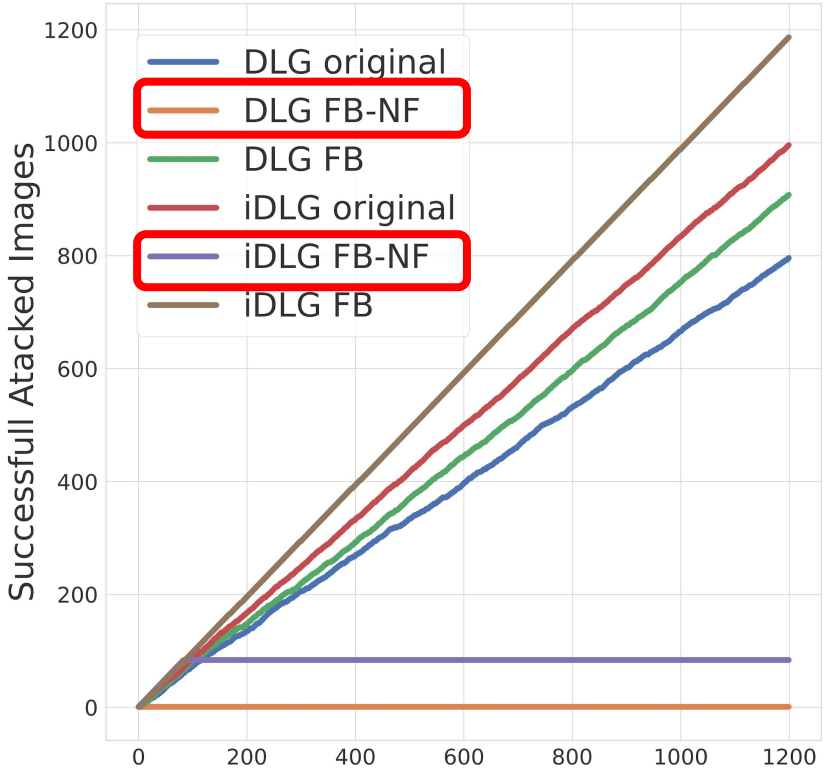


CIFAR100

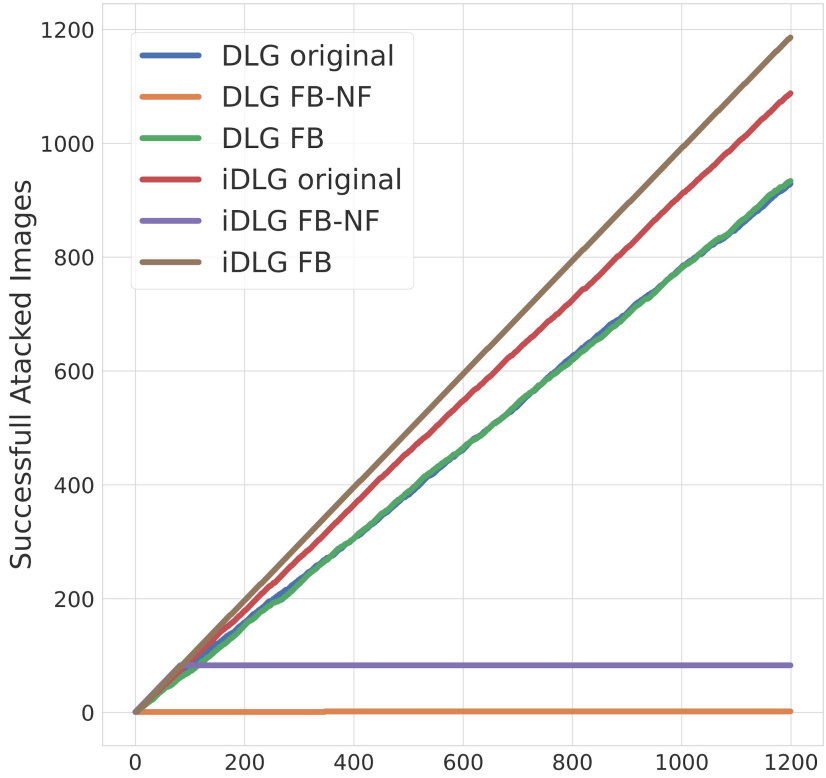


MNIST

Avaliação

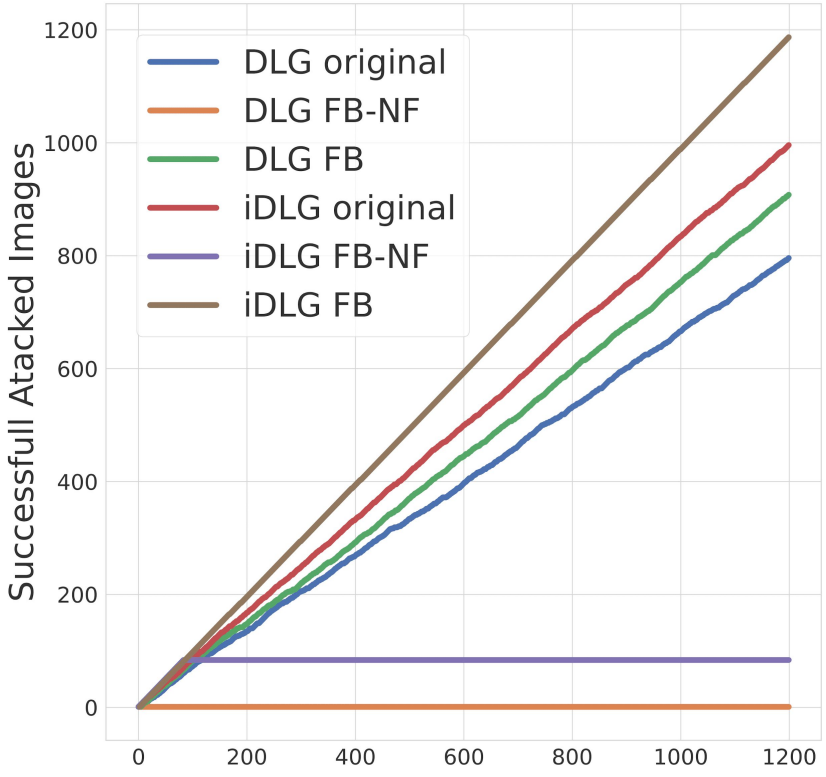


CIFAR100

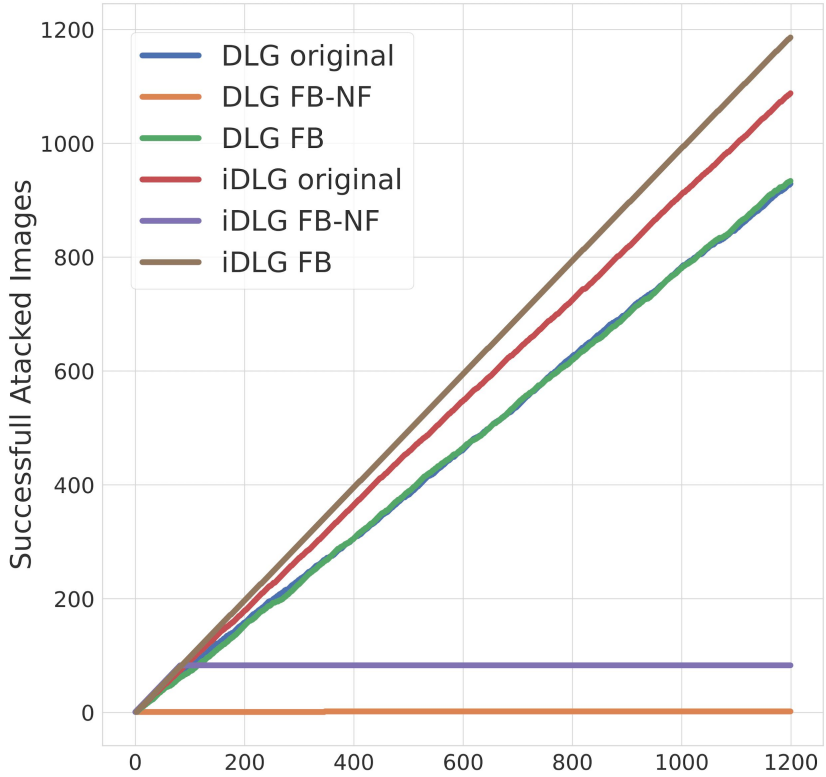


MNIST

Avaliação

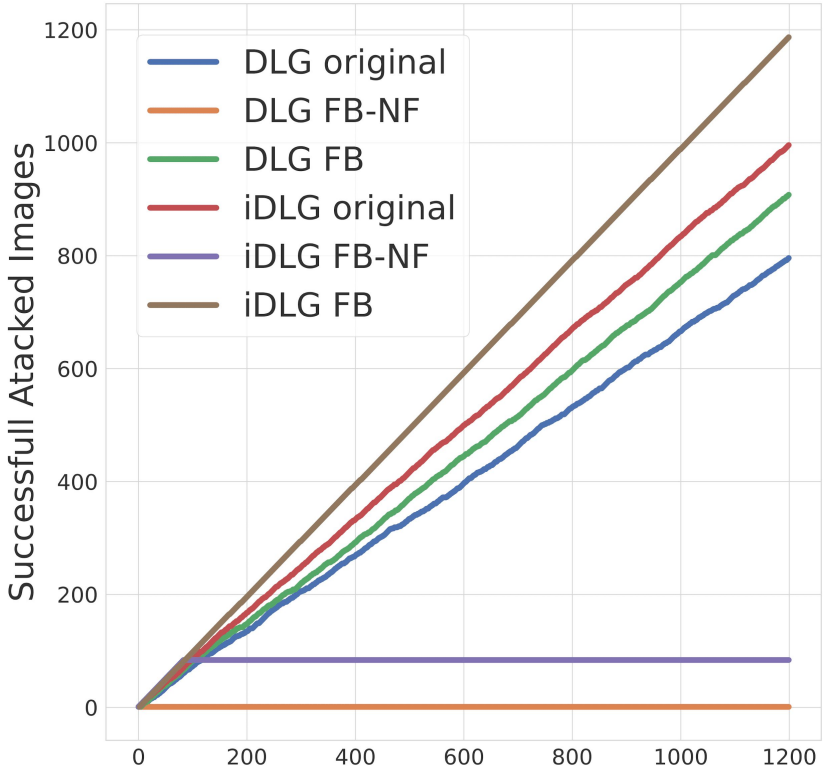


CIFAR100

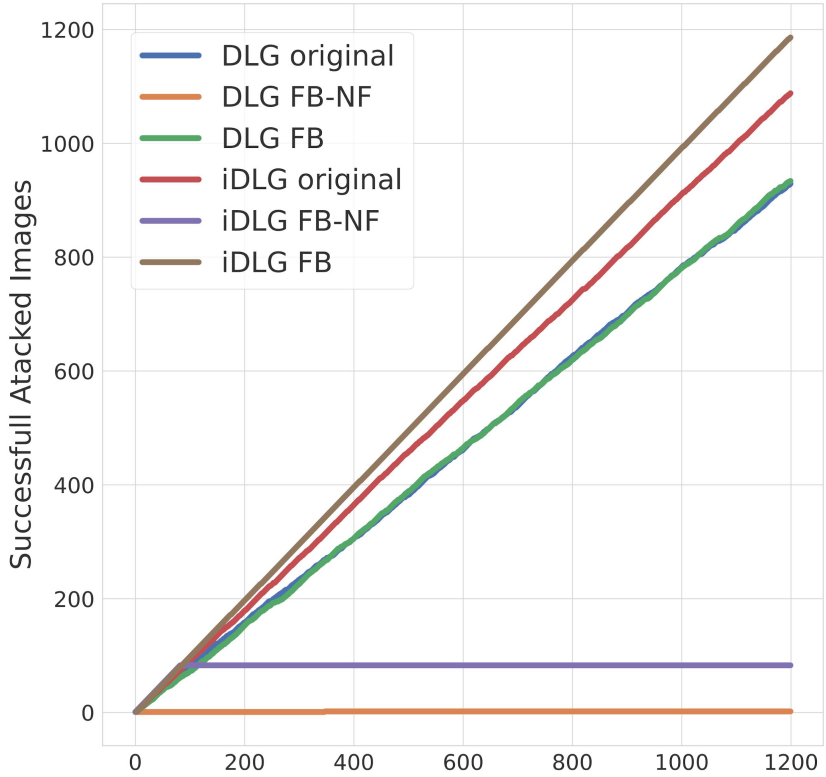


MNIST

Avaliação

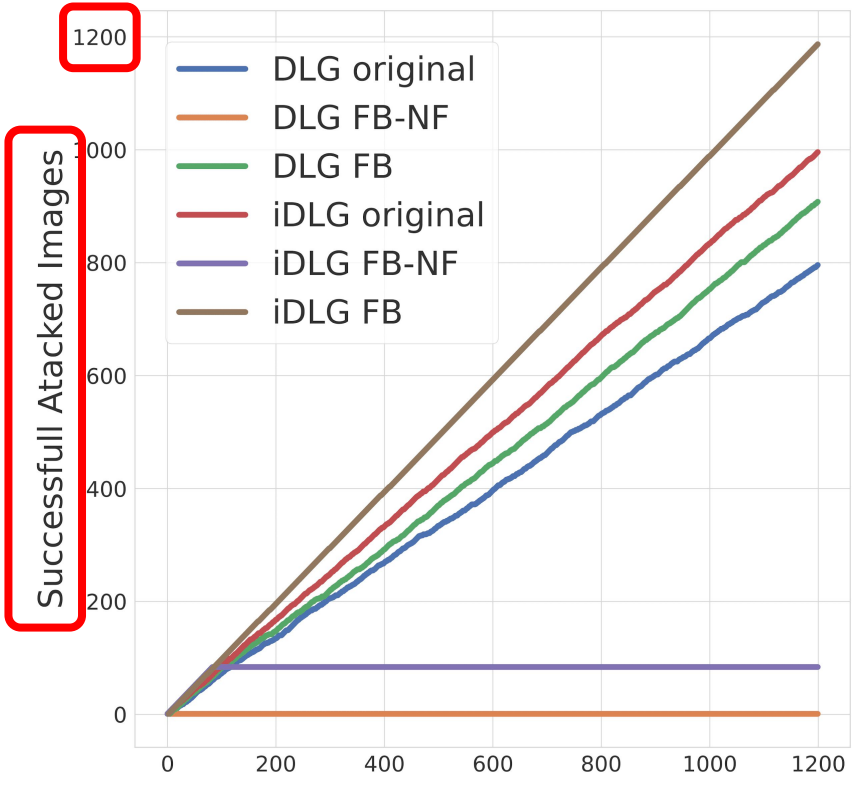


CIFAR100

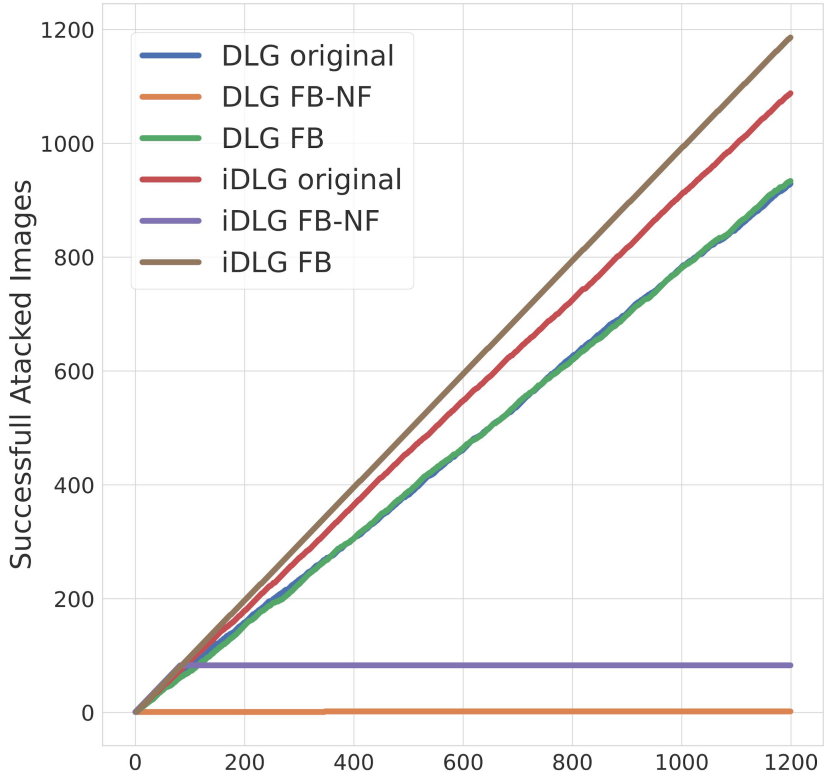


MNIST

Avaliação

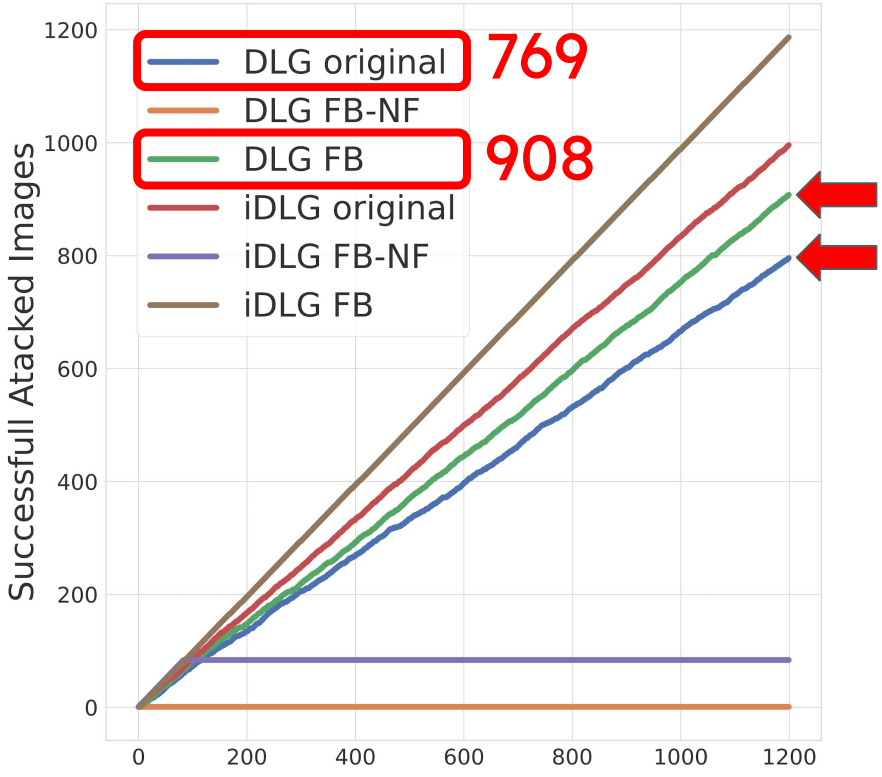


CIFAR100

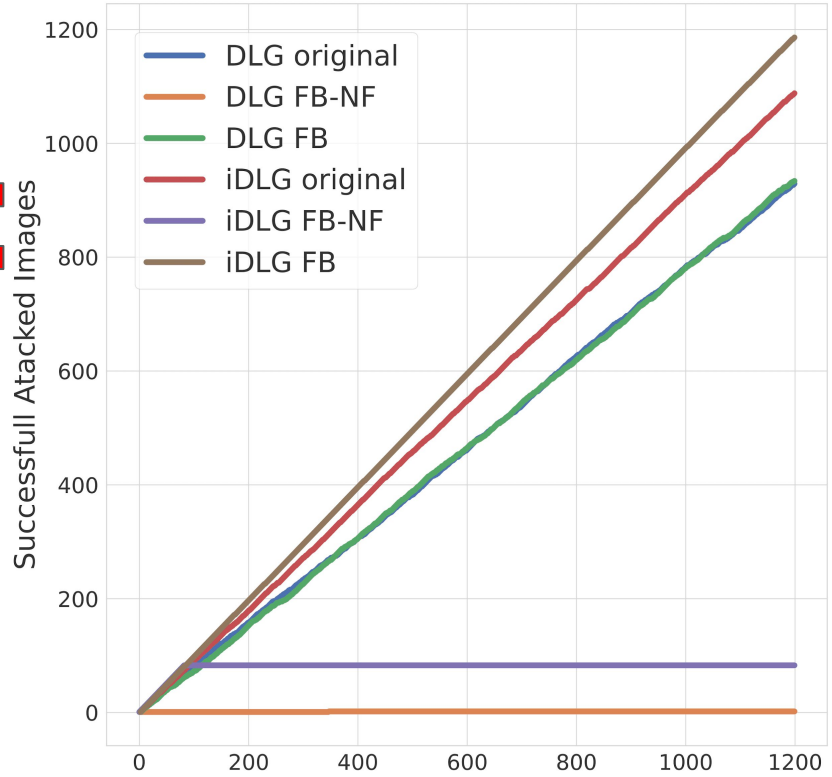


MNIST

Avaliação +14.07%

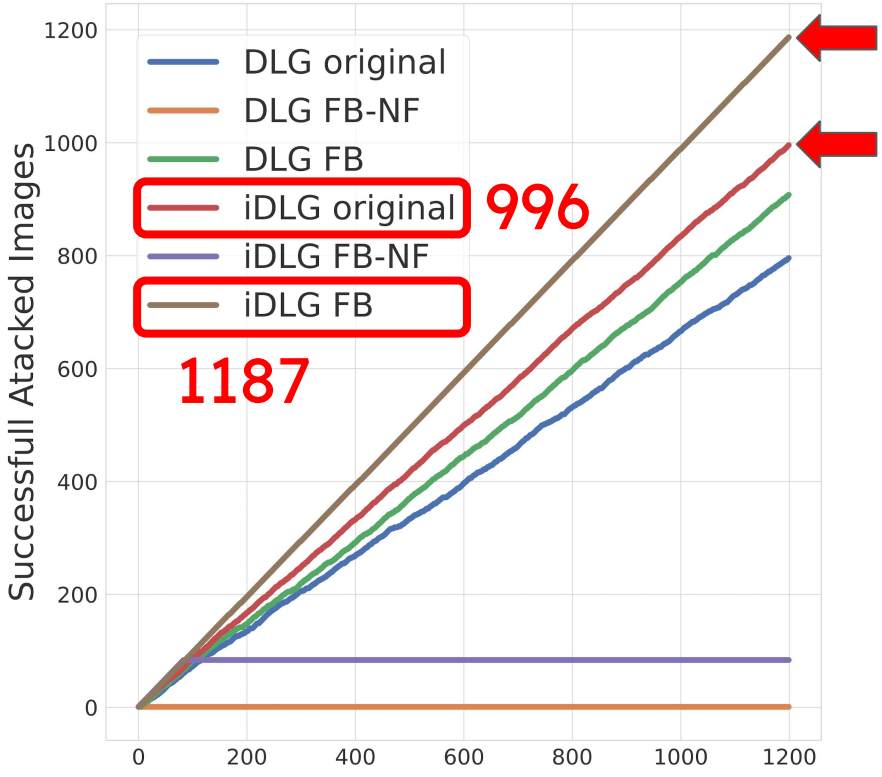


CIFAR100

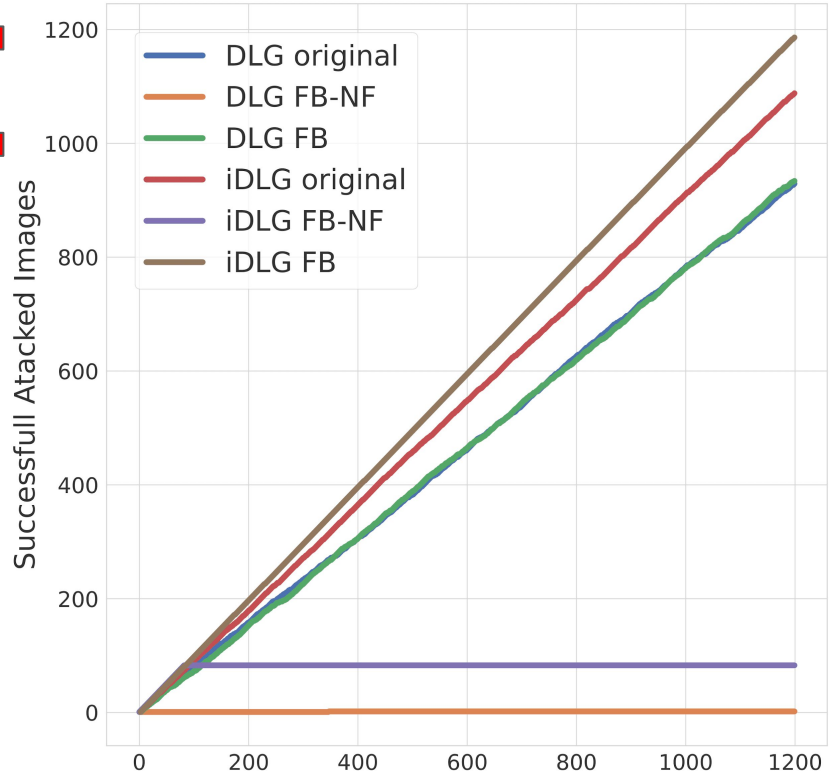


MNIST

Avaliação +19.18%

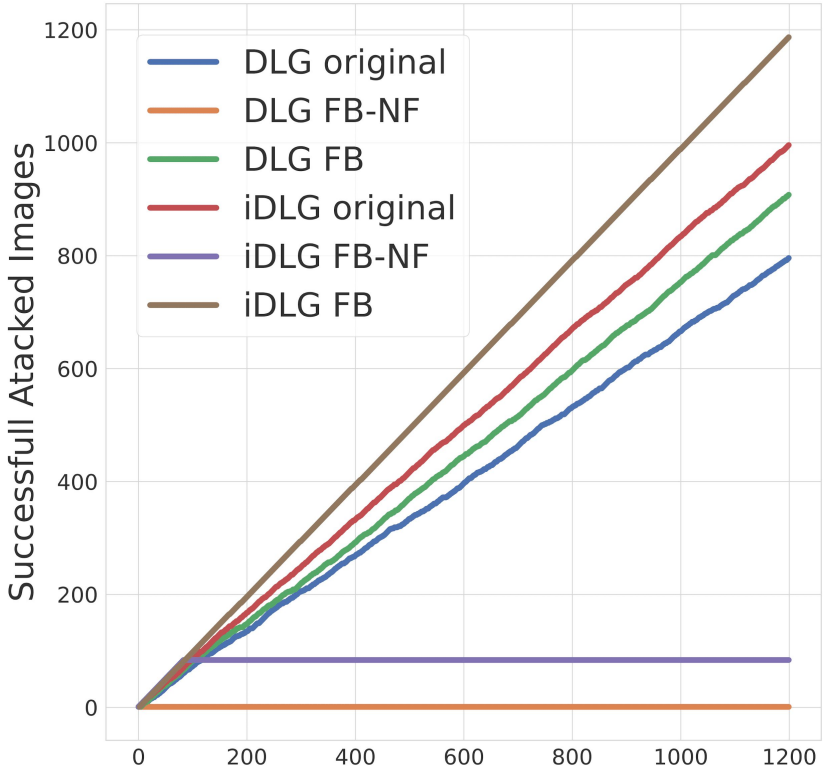


CIFAR100

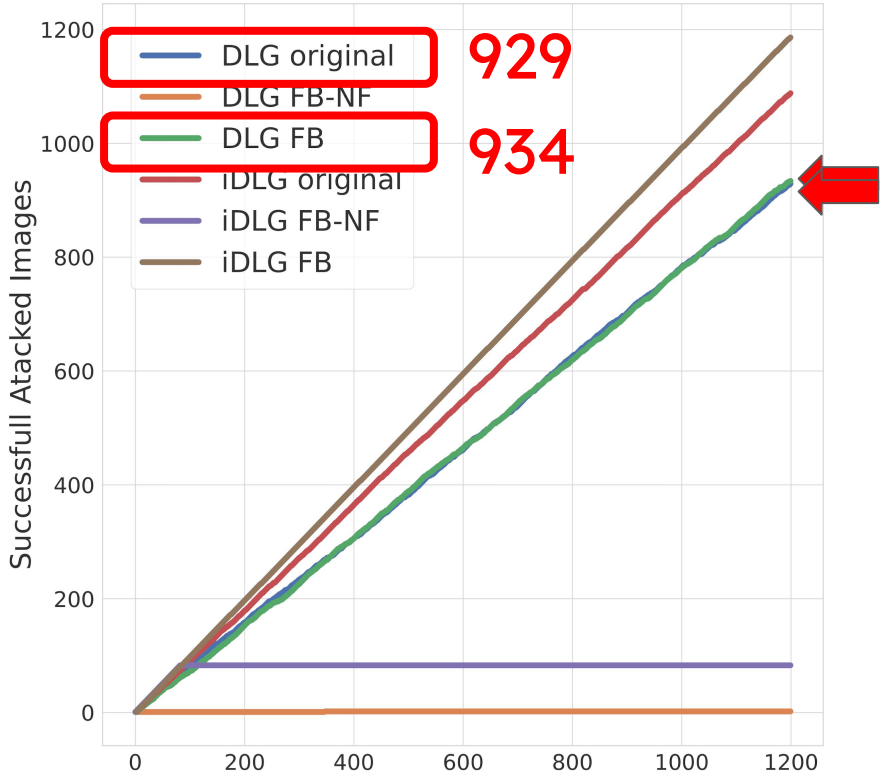


MNIST

Avaliação +5 imagens

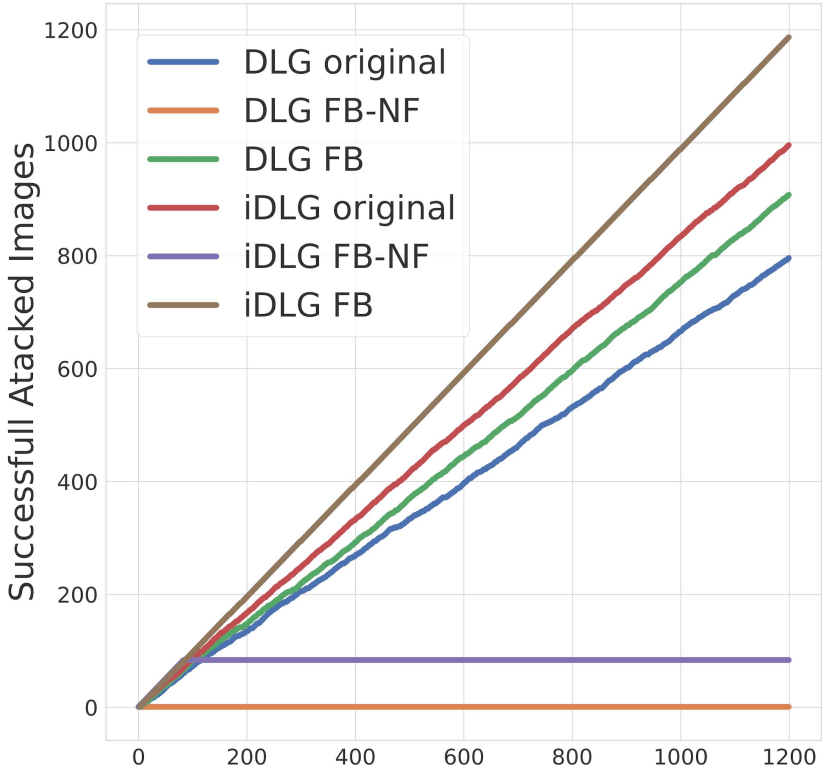


CIFAR100

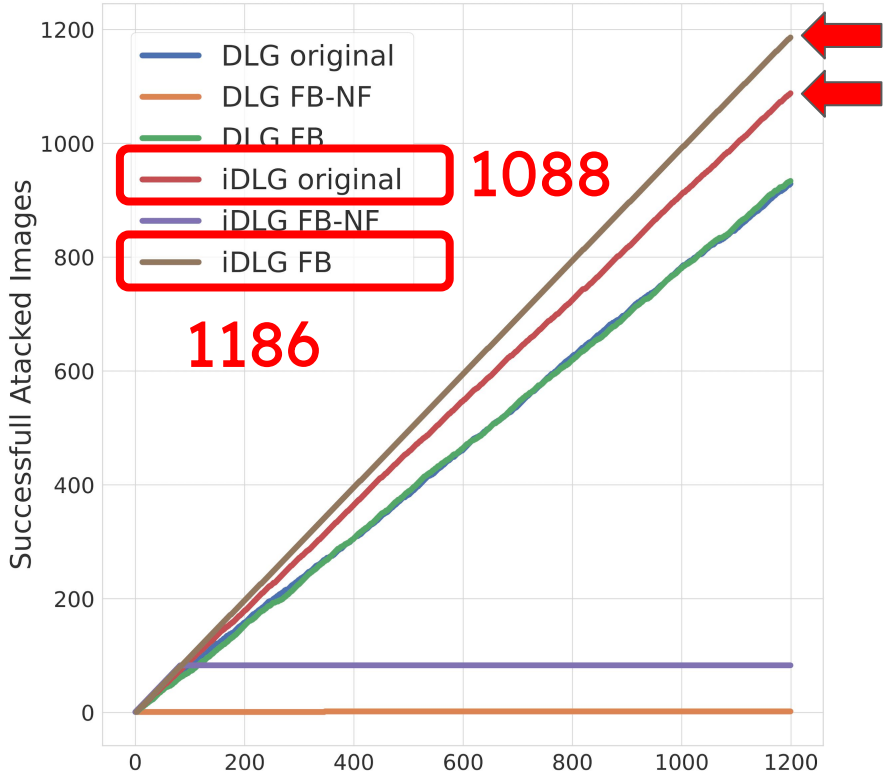


MNIST

Avaliação +9.01%

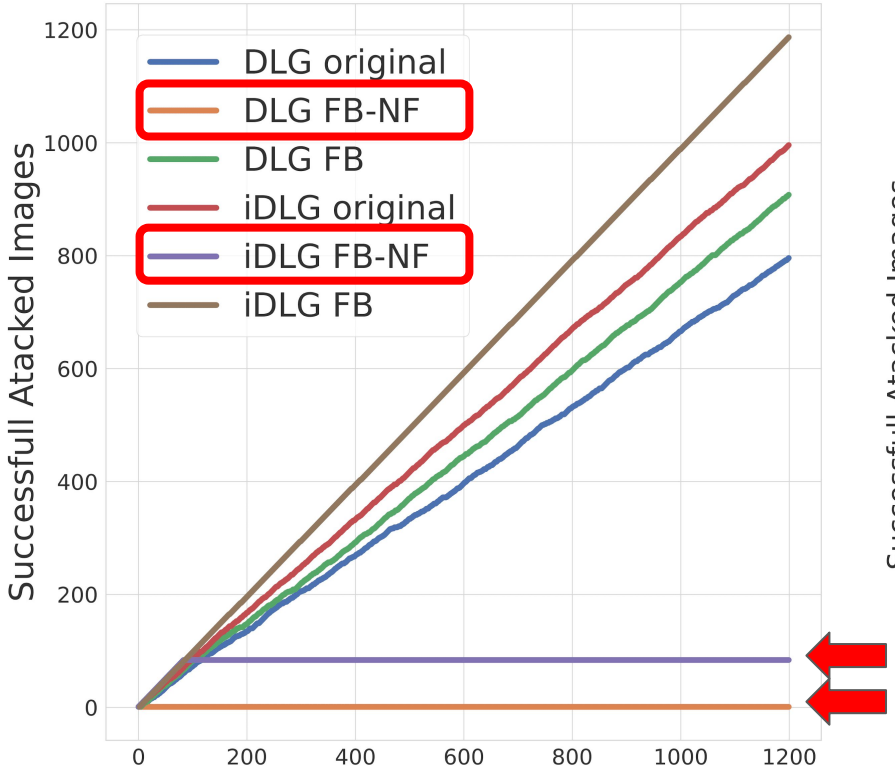


CIFAR100

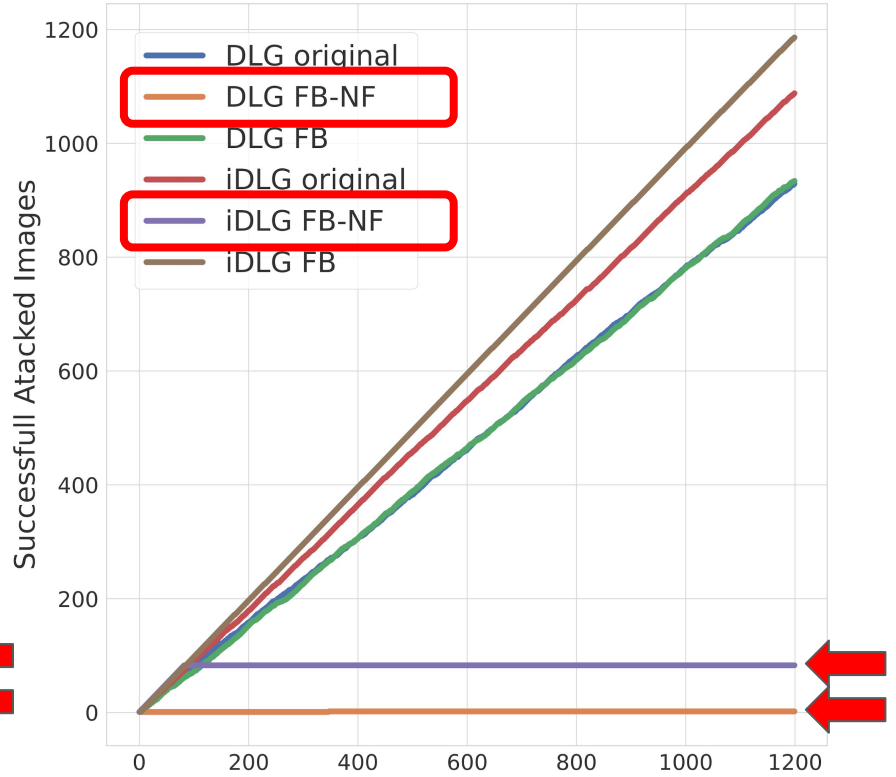


MNIST

Avaliação

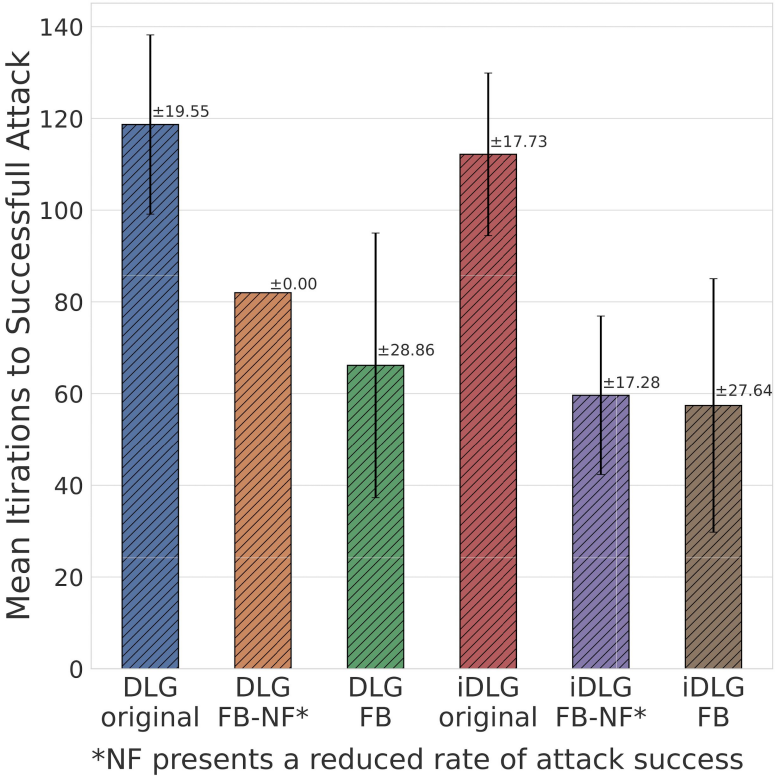


CIFAR100

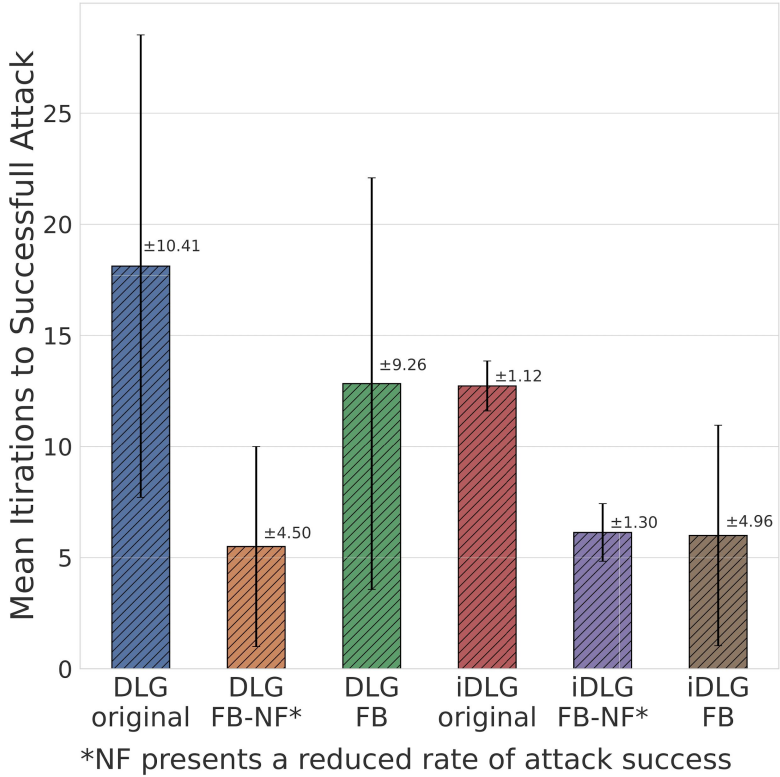


MNIST

Avaliação



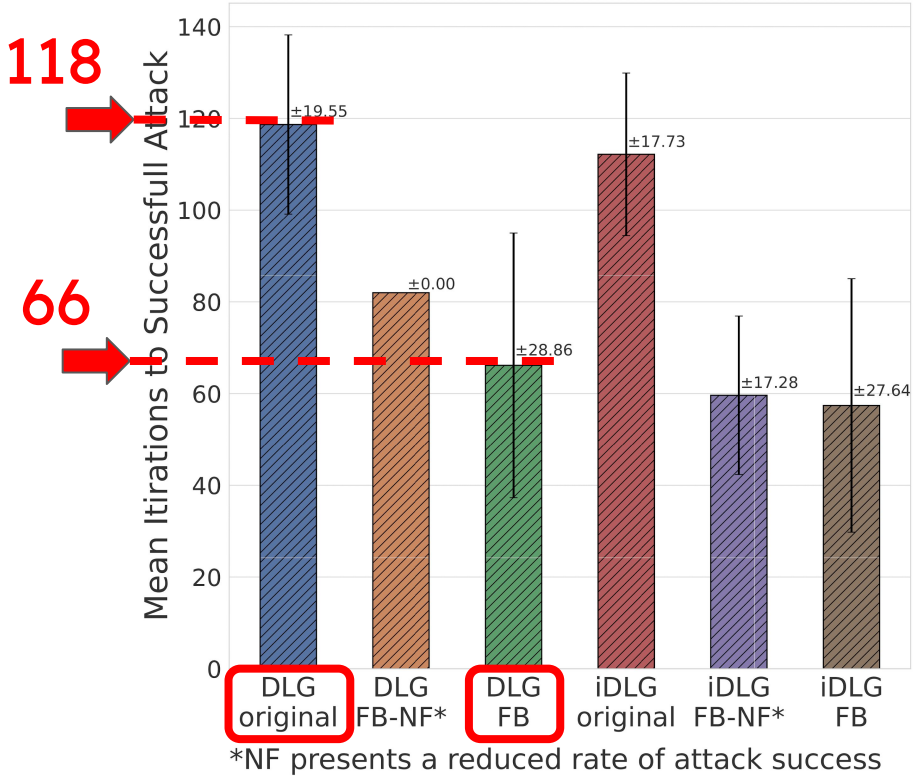
CIFAR100



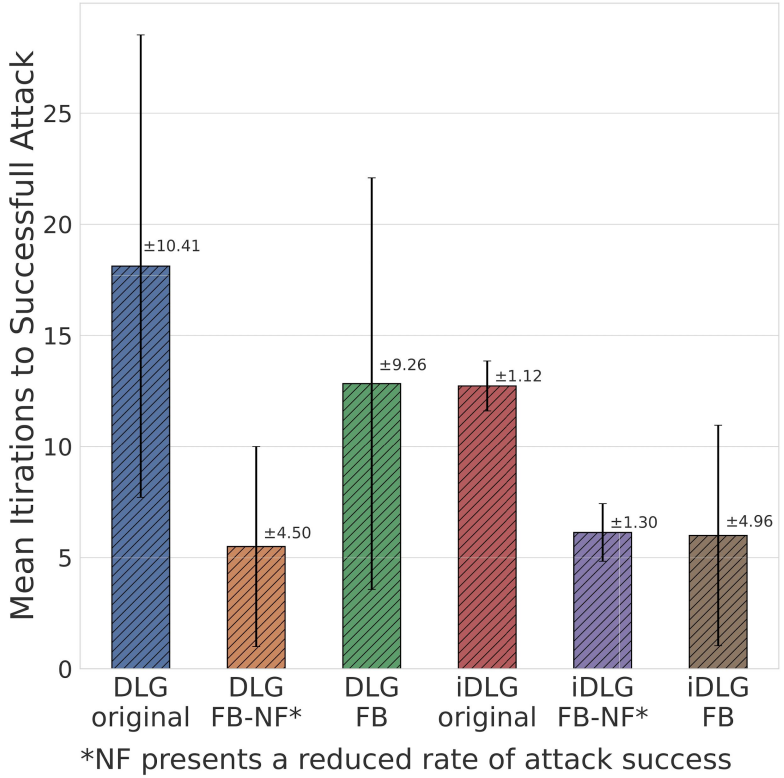
MNIST

SBSeg24 SÃO JOSÉ DOS CAMPOS

Avaliação 44,26%



CIFAR100

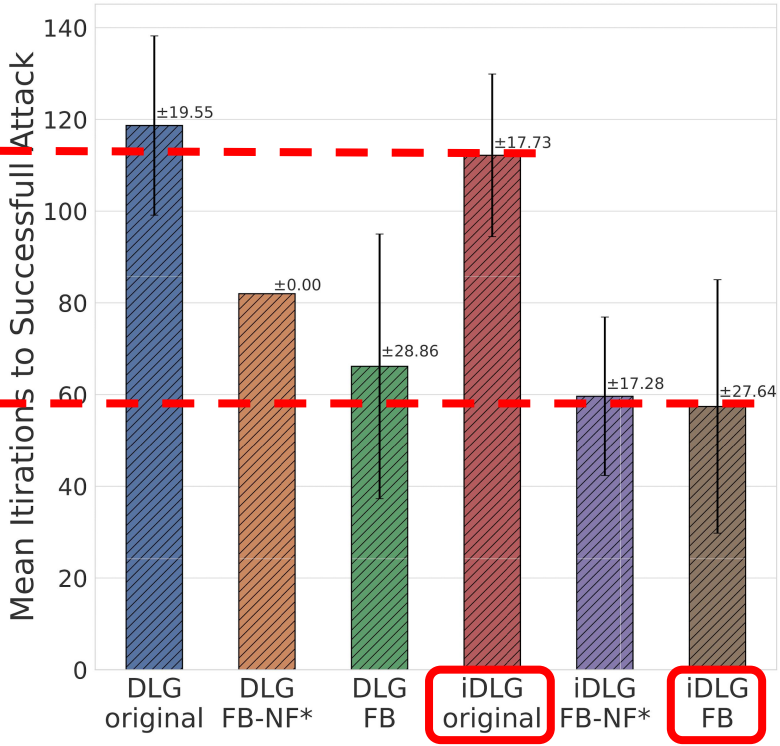


MNIST

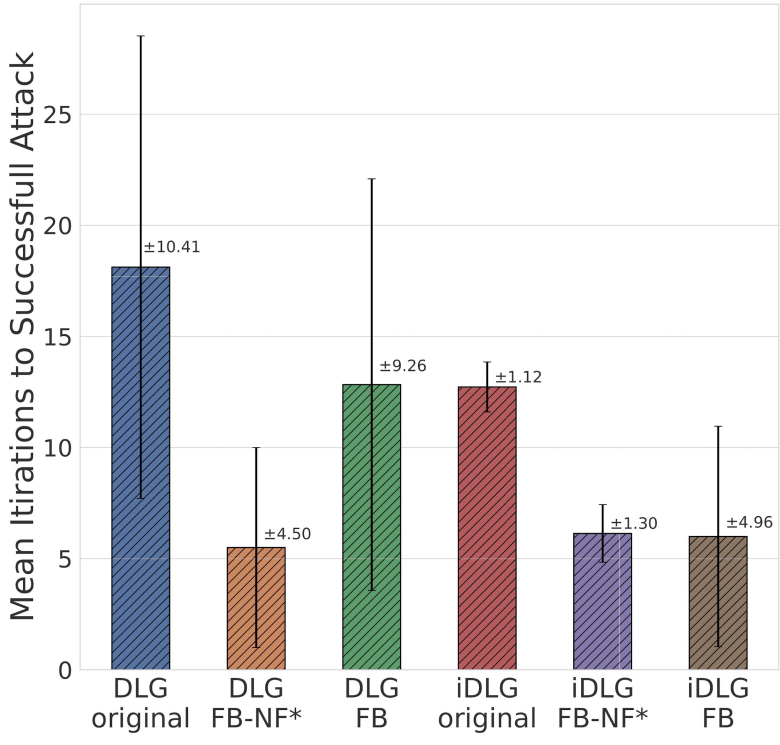
Avaliação 48,82%

112

57

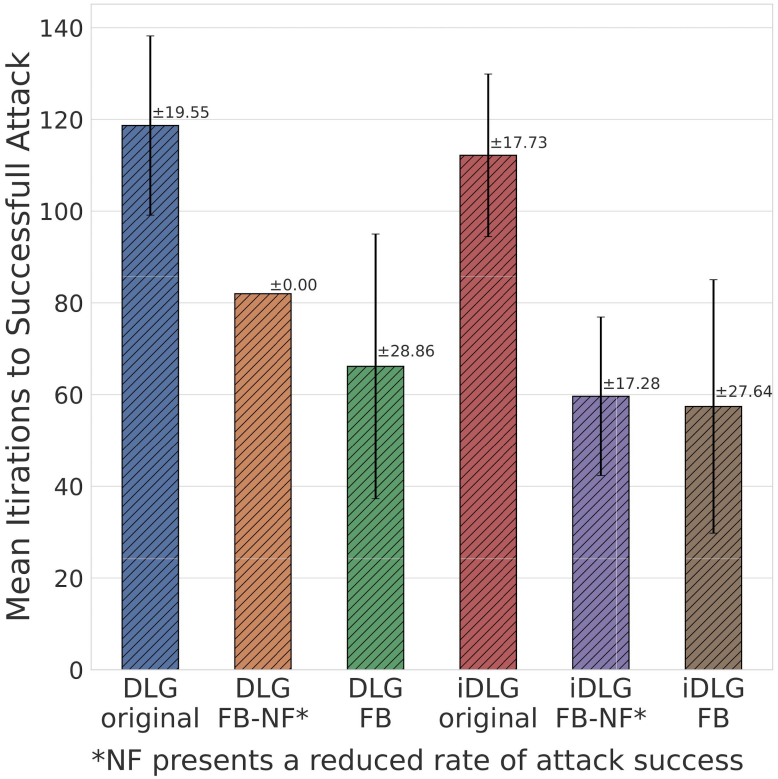


CIFAR100



MNIST

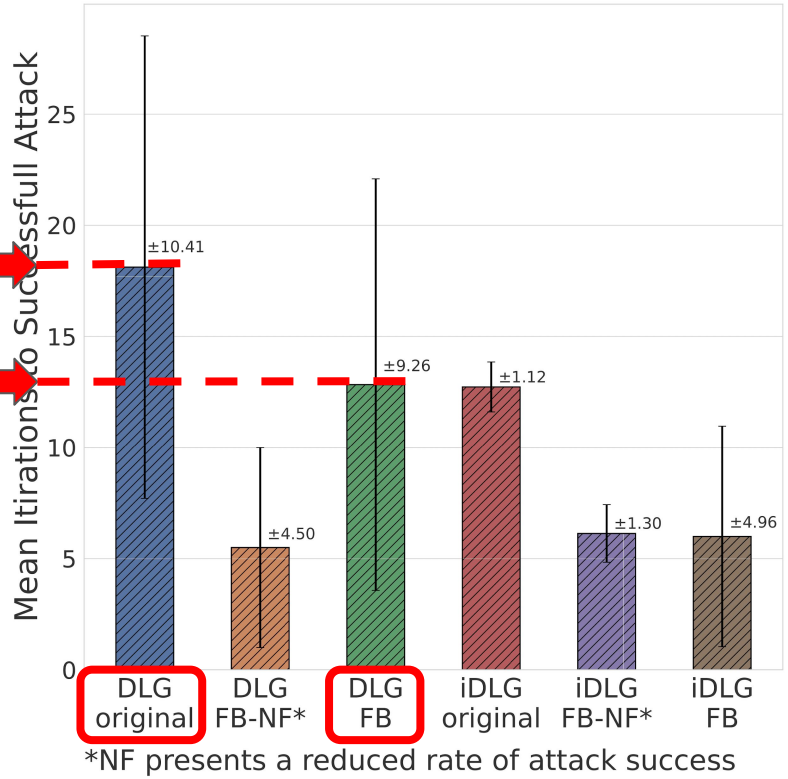
Avaliação 29,19%



CIFAR100

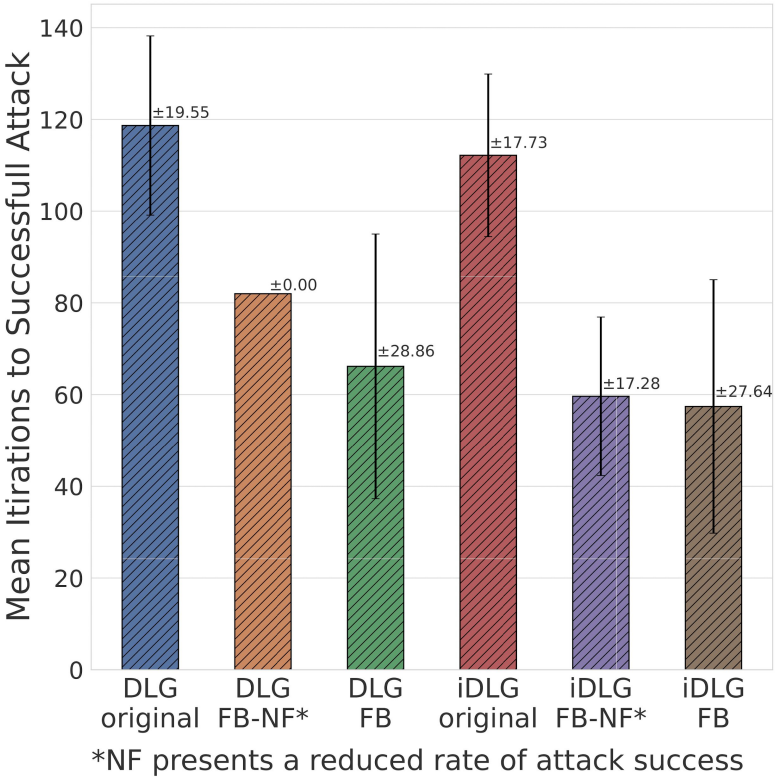
18

13

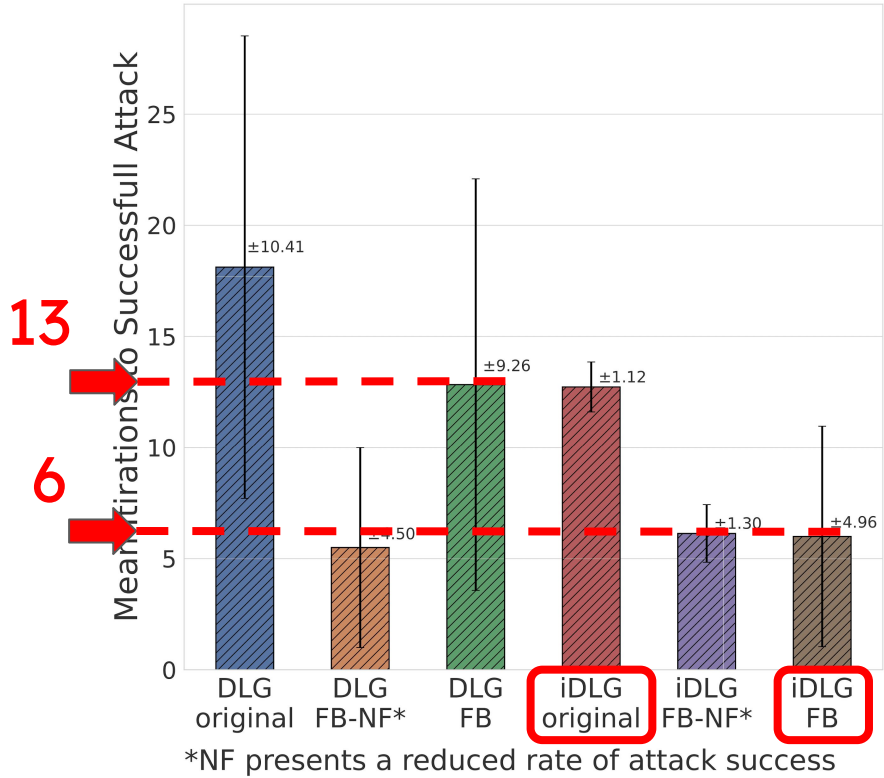


MNIST

Avaliação 52,59%



CIFAR100



MNIST

Considerações finais

- Sofisticação do ataque
  - +taxa de sucesso
  - -custo computacional
- Avanço contra 3 canais de cor

Trabalhos futuros

- Produzir mais métricas
- Analisar outros modelos
  - som, texto, vídeo
- Explorar defesas

Obrigado!

- Luiz Antônio Leite
  - luiz.freitas.leite@icen.ufpa.br
- Yuri Santo
- Prof. Bruno Dalmazo (FURG)
- Orientador:
  Prof. André Riker (UFPA)

Patrocinadores do SBSeg 2024!

nic.br

cgi.br

Google

Tempest

CAPES

SiDi

FAPESP

zscaler™

BugHunt

CNPq

C.E.S.A.R

FACULDADE IBPTECH