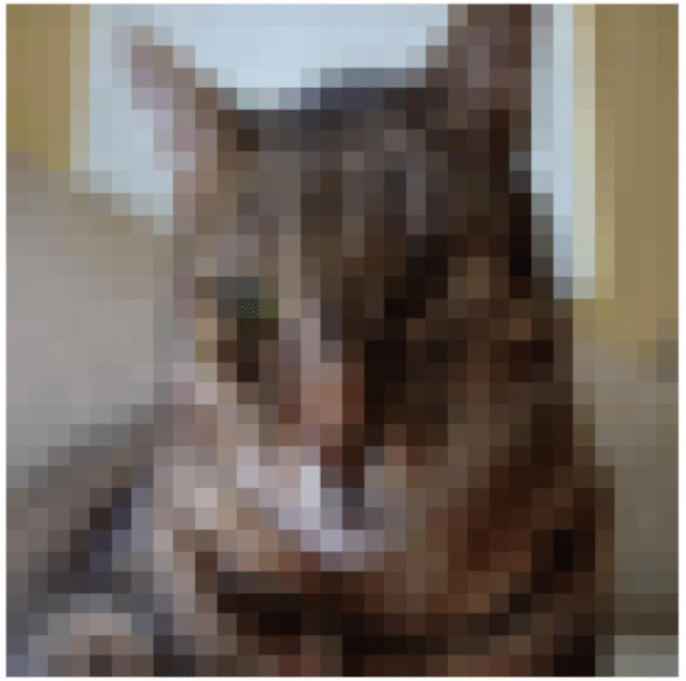
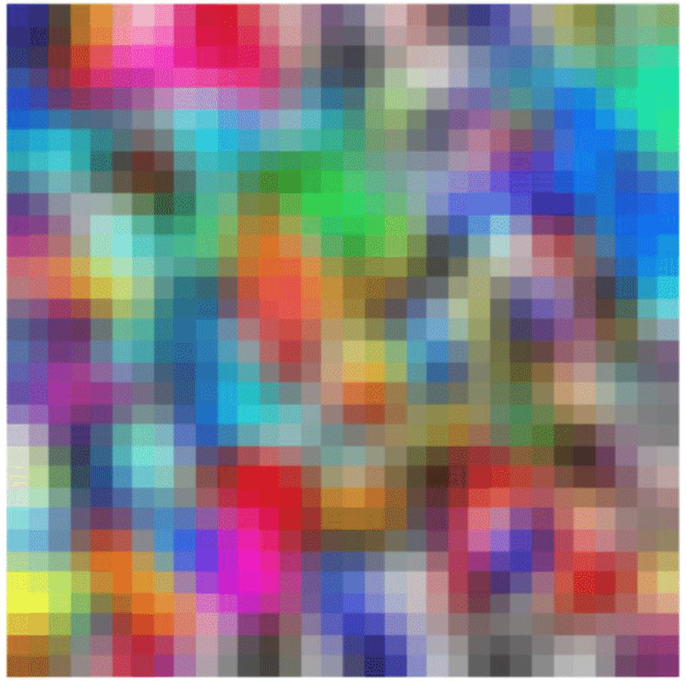
Original cat



$\boldsymbol{x}$

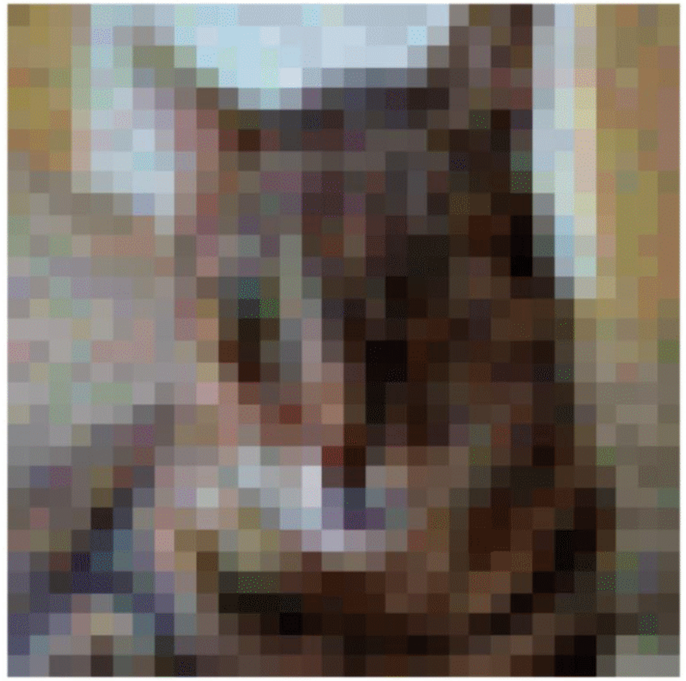
"cat"
94.6% confidence

Perturbation



$0.031\times$
$\boldsymbol{sign}(\boldsymbol{\nabla}_{\boldsymbol{x}}\ell(\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y}))$

Adversarial sample



$\boldsymbol{x} +$
$\boldsymbol{sign}(\boldsymbol{\nabla}_{\boldsymbol{x}}\ell(\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y}))$

"dog"
99.8% confidence