# GenAttack: Practical Black-box Attacks with Gradient-Free Optimization

**Moustafa Alzantot** (UCLA), **Yash Sharma**(Cooper Union), **Supriyo Chakraborty (**IBM Research), **Mani Srivastava** (UCLA)

## Practical black box attacks

- Machine learning models can be fooled by small but maliciously crafted perturbations of “adversarial attacks”.
- White box attackers utilize their knowledge of model weights and architecture to compute the exact network gradients to attack a model.
- In a practical threat model, the attacker does not have access to this information.
- Previous research in blackbox attack is based on either:
  - *Training substitute models*
  - *Zeroth order optimization that approximates the gradient by applying difference equations to the results of model queries.*

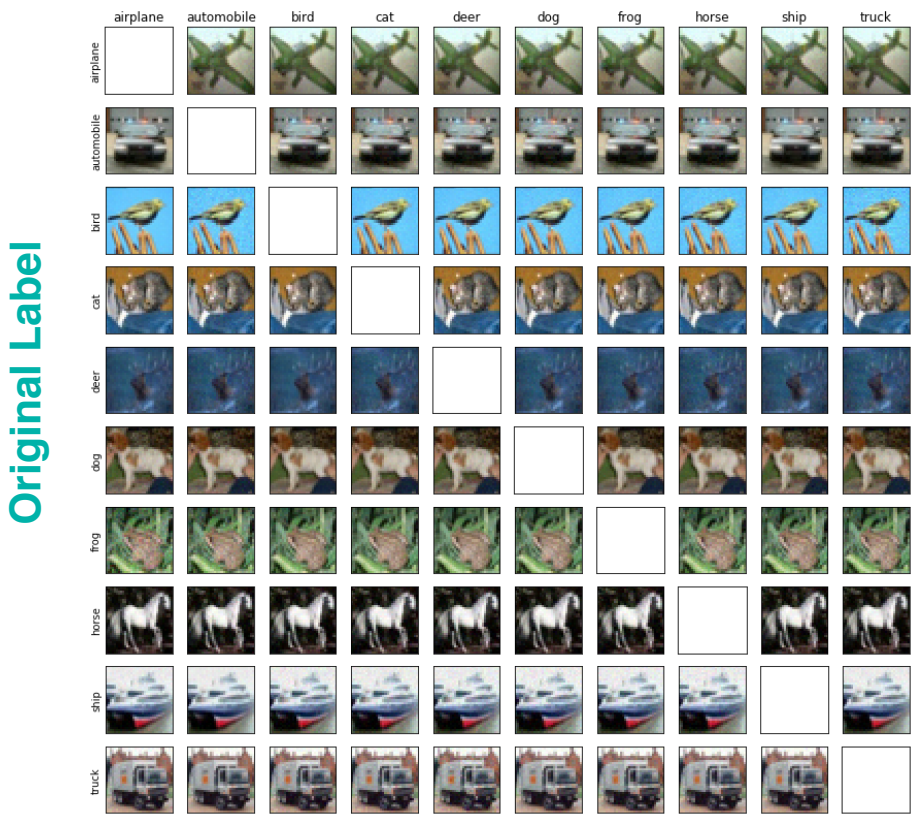
Both approaches are query inefficient and suffer from low attack success rate

## GenAttack

### Gradient Free Optimization Approach

- We propose a gradient-free based approach to generate adversarial examples.
- Following the biologically-inspired genetic optimization approach:
  - Initial generation of candidate attacks computed by adding small random noise.
  - Evaluate fitness scores of population members.
  - Select the strongest population members. Apply crossover and mutation to find next generation.

## GenAttack Results CIFAR-10




## Inception-v3

Karatoe Galerita

Trolly bus

## Query efficiency

| | ZOO | GenAttack |
|---|---|---|
| MNIST | 2,118,222 | 996 **(2,126X)** |
| CIFAR-10 | 2,064,798 | 804 **(2,568X)** |
| ImageNet | 2,611,456 | 97,493 **(27X)** |

**Distributed Analytics and Information Science International Technology Alliance**
Annual Fall Meeting, September 2018