

# How to break Apple's NeuralHash

Introduction to adversarial preimage attacks



sogeti  
Part of Capgemini

h25

# Apple's PSI protocol

- Used to detect CSAM on iCloud
- Verification is on-device only
- Privacy-preserving crypto

# Apple's PSI protocol

- Detection is based on a watchlist of known hashes
- Standard hashing is inefficient (easy to avoid detection by flipping any bit)  
→ use Neural Networks!

# NeuralHash

Neural network trained to give the same hash to similar images

Very similar to FADA's FaceNet!

# How it's trained

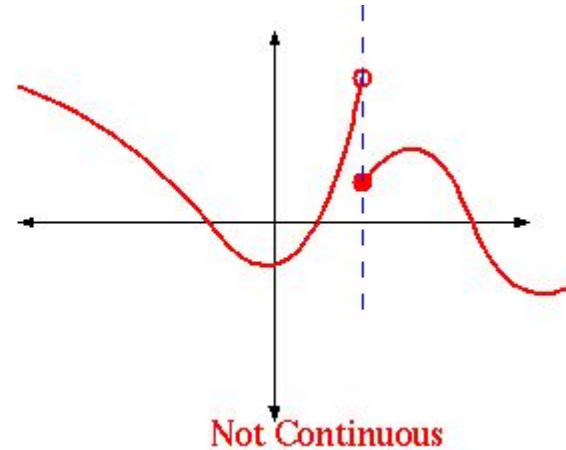
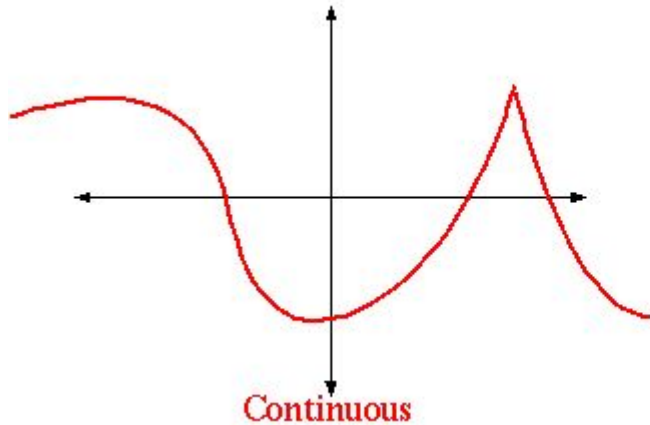


**Corporate needs you to find the differences between this picture and this picture.**

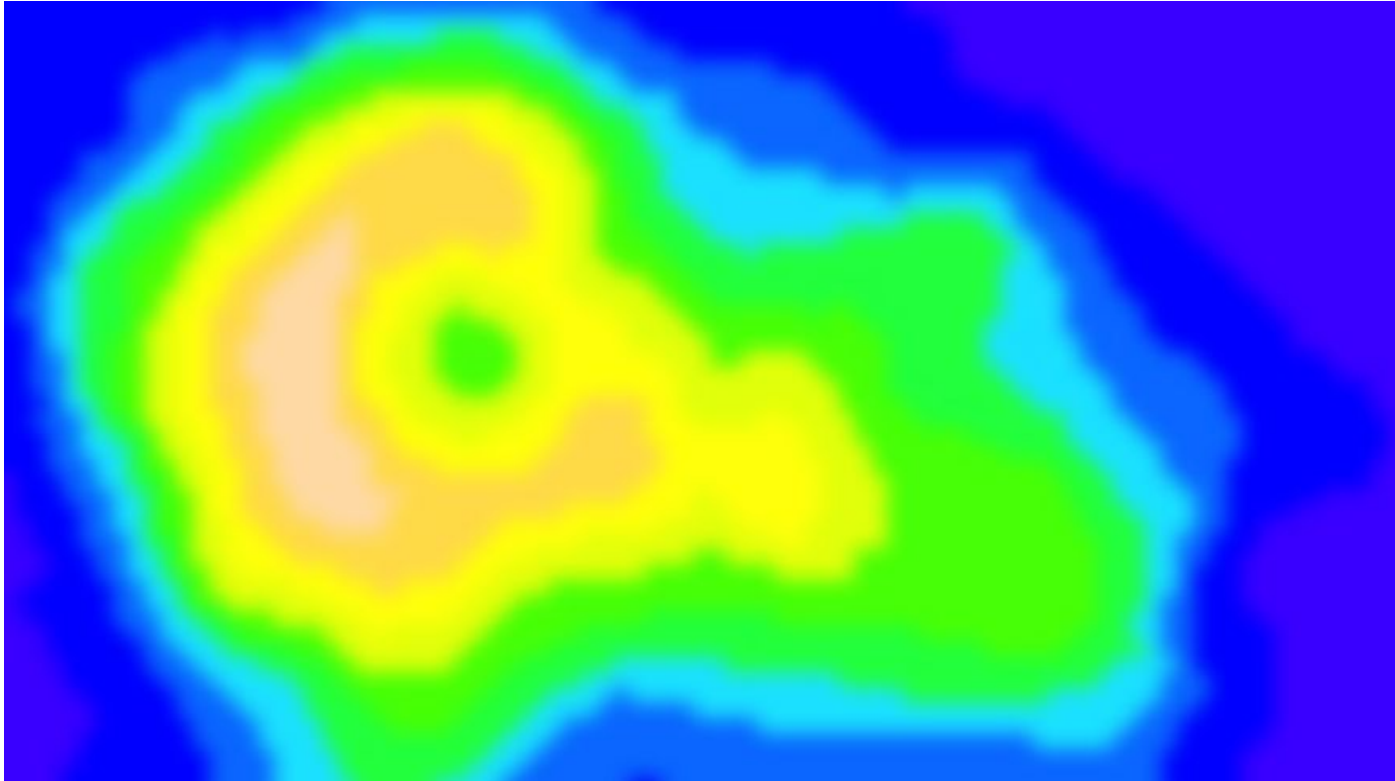
# Adversarial attacks

Neural networks are continuous functions

Small input change = small output change



# Adversarial attacks



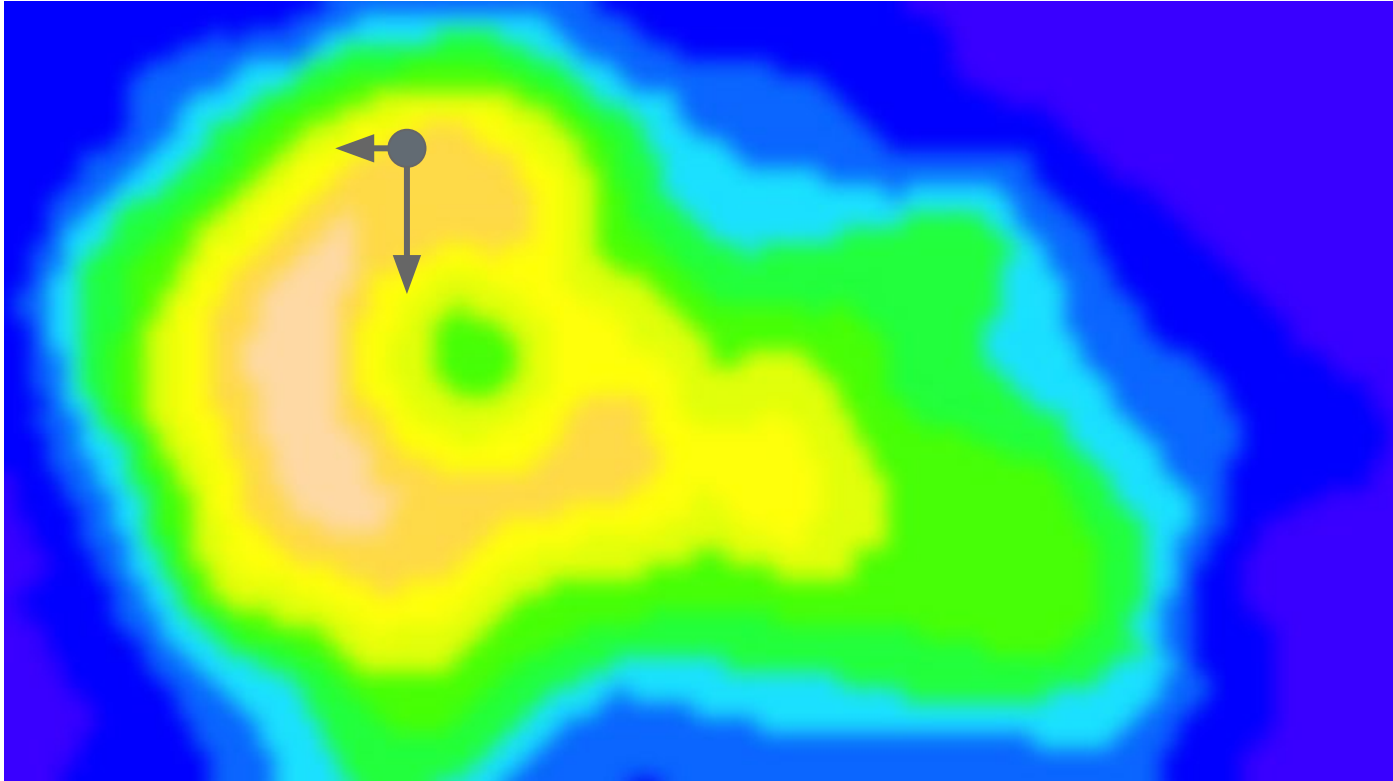
# Adversarial attacks

## Differentiability

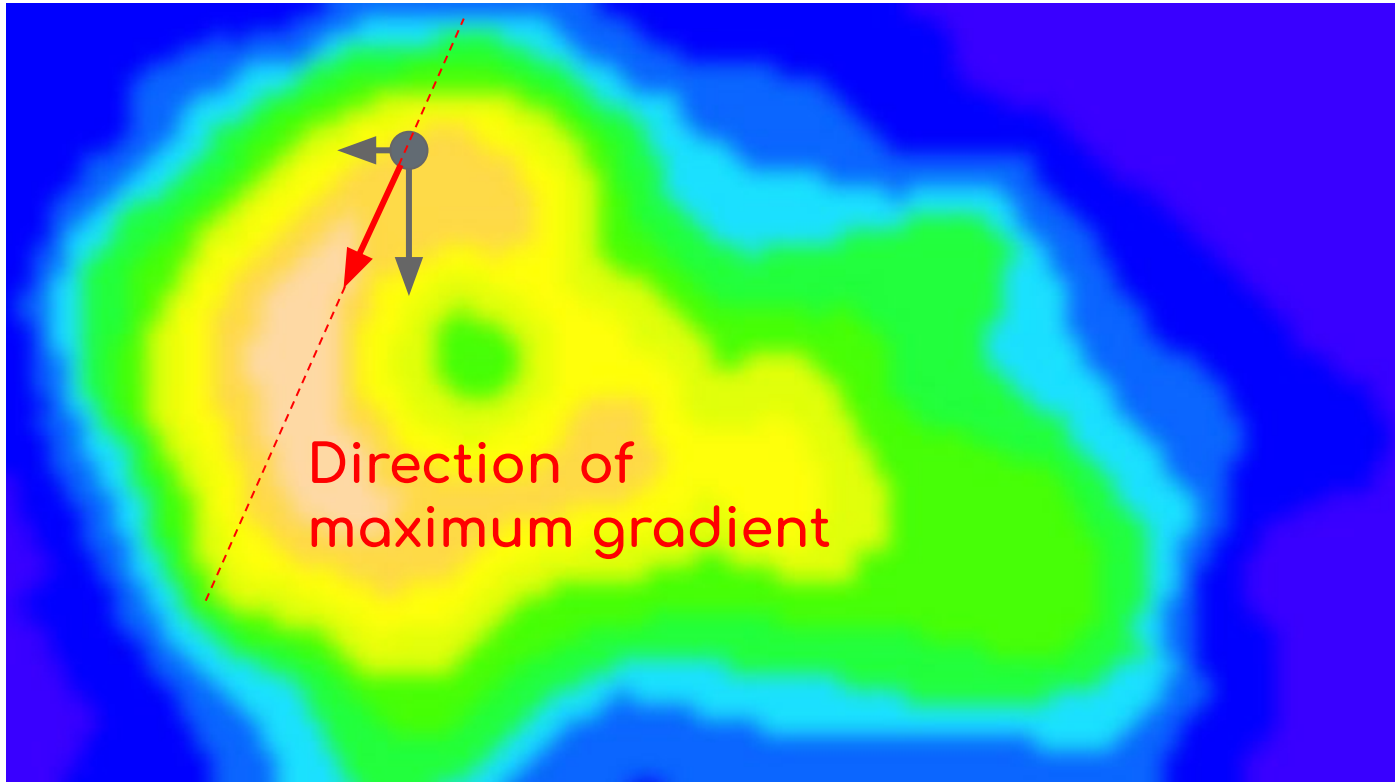
is a crucial property of neural networks



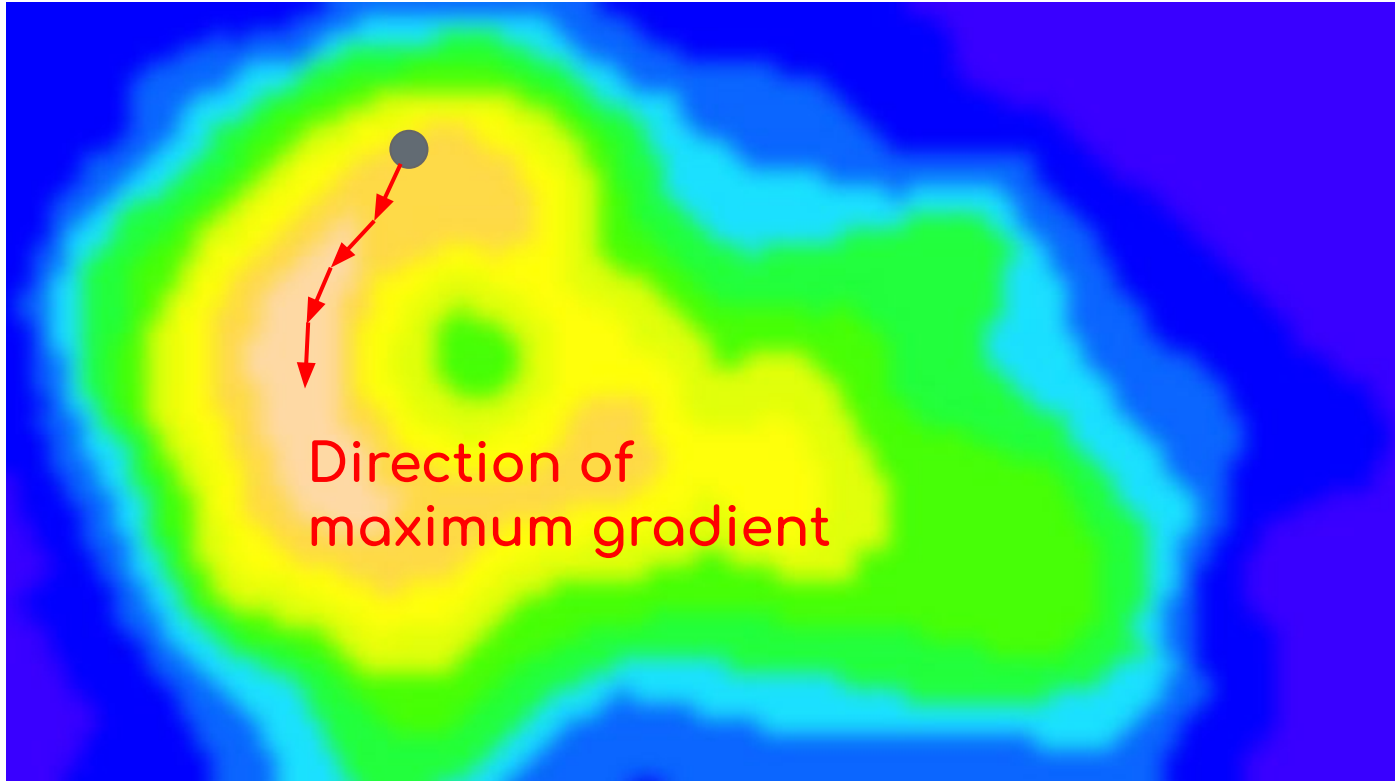
# Adversarial attacks



# Adversarial attacks



# Adversarial attacks



# Adversarial attacks

Much more complex in practice (1080p image is more than 6M dimensions)

But the principle is the same : small changes towards target using gradient

Let's hack NeuralHash!

# Result



b8c8cd71e551101f54c3874b

# Conclusion

- All neural networks have this vulnerability by definition
- Some mitigations exist
- Do not use NNs for critical tasks in adversarial environments

# Bonus





# Bonus



= c7ef b04a f72a...

# Bonus



= c7ef b04a f72a...

# Bonus



= c7ef b04a f72a...



= c7ef b04a f72a...

Q&A

Thanks!



@MathisHammel / @h25io



discord.h25.io



twitch.h25.io

