

ADVANCE ALL MARCH EVERLASTING

TMELINE











2

0

0

4

2

0

1

3

A pixelated, black and white representation of the number 2. The digit is composed of a thick, blocky stroke with a curved top and a horizontal base. The edges are jagged, characteristic of low-resolution digital art.A pixelated, black and white representation of the number 0. The digit is a thick, blocky oval shape with a horizontal base. The edges are jagged, characteristic of low-resolution digital art.A pixelated, black and white representation of the number 1. The digit is a thick, blocky vertical stroke with a horizontal base and a short horizontal bar at the top. The edges are jagged, characteristic of low-resolution digital art.A pixelated, black and white representation of the number 4. The digit is a thick, blocky shape with a vertical stem, a horizontal crossbar, and a diagonal stroke on the left. The edges are jagged, characteristic of low-resolution digital art.



2

0

1

5



2

0

1

6

A red speech bubble border with a pointed bottom, containing the text "FOOLING THE SPAM FILTER".

**FOOLING THE
SPAM FILTER**



**FIRST ATTEMPT TO
FOOL A NEURAL
NET**

IMPERCEPTIBLE ADVERSARIAL EXAMPLES



20

0

1

7



ADVERSARIAL TRAINING AS DEFENCE



DEFENSIVE DISTILLATION

A red speech bubble border with a pointed bottom, containing the text.

ADVERSARIAL EXAMPLES IN PHYSICAL WORLD & BLACK BOX ATTACKS



LARGE SCALE ADVERSARIAL TRAINING & ENSEMBLE APPROACHES

ADVERSARIAL 3D MODELS & STRONGER L2 ATTACKS

A red outline of a speech bubble with a pointed bottom, containing the text "FAST & PRACTICAL ATTACK MODEL".

FAST & PRACTICAL ATTACK MODEL

HISTORY OF ADVERSARIAL MACHINE LEARNING

- ▶ Adversarial Classification by Dalvi et al. 2004 (Fooled Spam Filter)
- ▶ Evasion Attacks Against Machine Learning at Test Time by B. Biggio et al. (Fooled Neural Net)
- ▶ Intriguing properties of neural networks by C. Szegedy et al. 2014 (Imperceptible Adversarial Examples)
- ▶ Explaining and Harnessing Adversarial Examples by Goodfellow et al. 2015 (Fast Attack Model)
- ▶ Adversarial Examples in Physical World by Kurakin et al. 2016 (Demonstrates Adversary in Real World)
- ▶ Adversarial Machine Learning at Scale by Kurakin et al. 2017 (Adversarial Training)
- ▶ Practical Black Box Attacks by Papernot et al. 2016 (Practical Black Box Attacks)
- ▶ Evaluating the robustness of neural networks by Wagner et. al. 2017 (Strong L2 Attack)
- ▶ Ensemble approach for adversarial defence by Papernot et al. 2017 (Ensemble Defence)

TIMELINE

