



# 研究概要

Akinori Minagi    Hokuto Hirano  
Kazuhiro Takemoto

# 目的：

AI を間違えさせる。

AI の信頼性を評価する。

# 実験環境：

OS: Ubuntu 18.04.3 LTS

Python: v3.7.0

FrameWork:

tensorflow v1.14.0

keras v2.2.4

# -- LINKS --

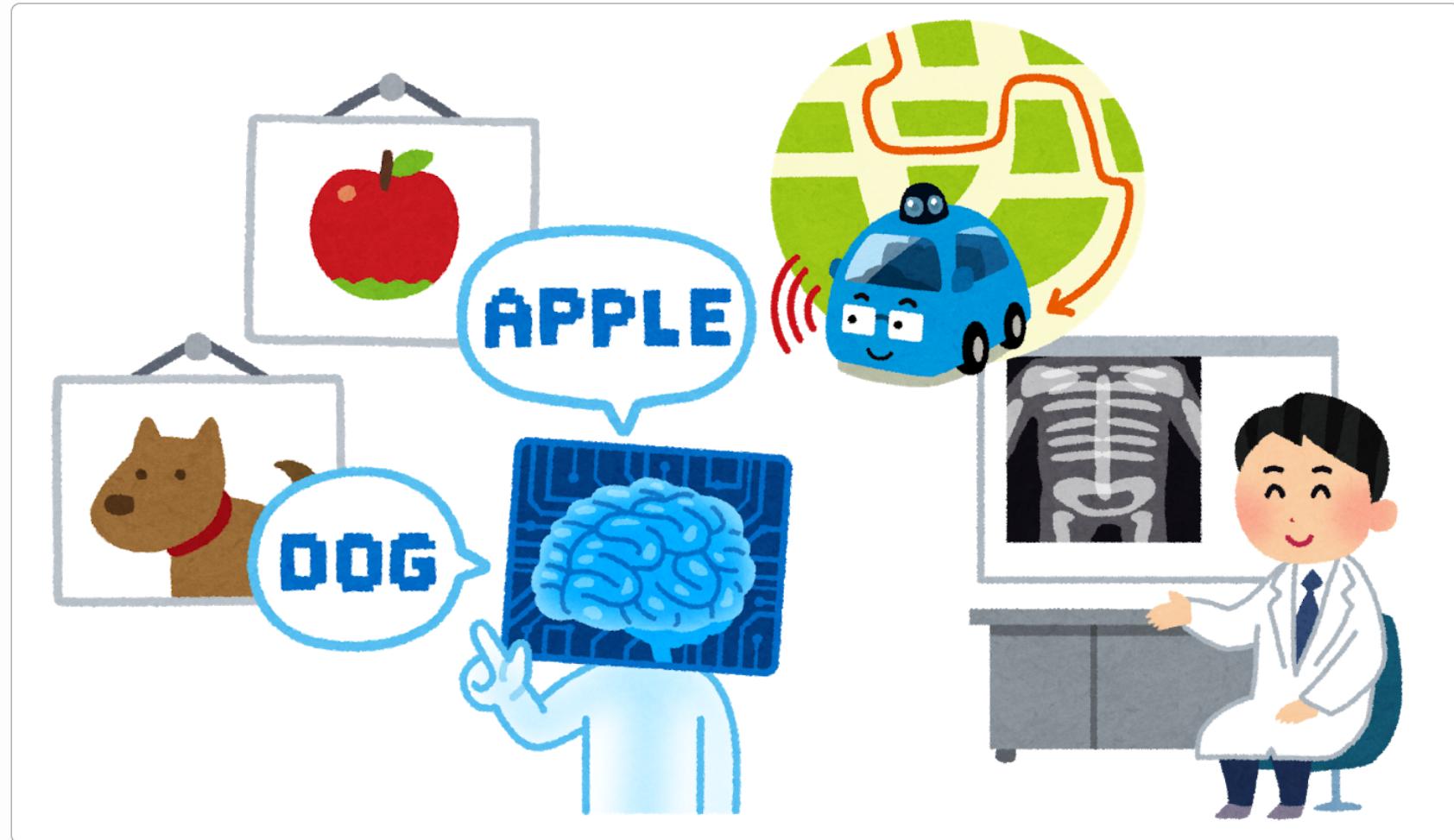
GitHub:

<https://github.com/Git-Gimi-Git/UAP>

Google Drive:

[https://drive.google.com/drive/folders/1mrS1hJZKchQIfRTrsQEoyXEjGaHiXoY\\_?usp=sharing](https://drive.google.com/drive/folders/1mrS1hJZKchQIfRTrsQEoyXEjGaHiXoY_?usp=sharing)

# AIのいま



AIはもうすぐ、そばに。

AIは敵対的擾動で、間違える。

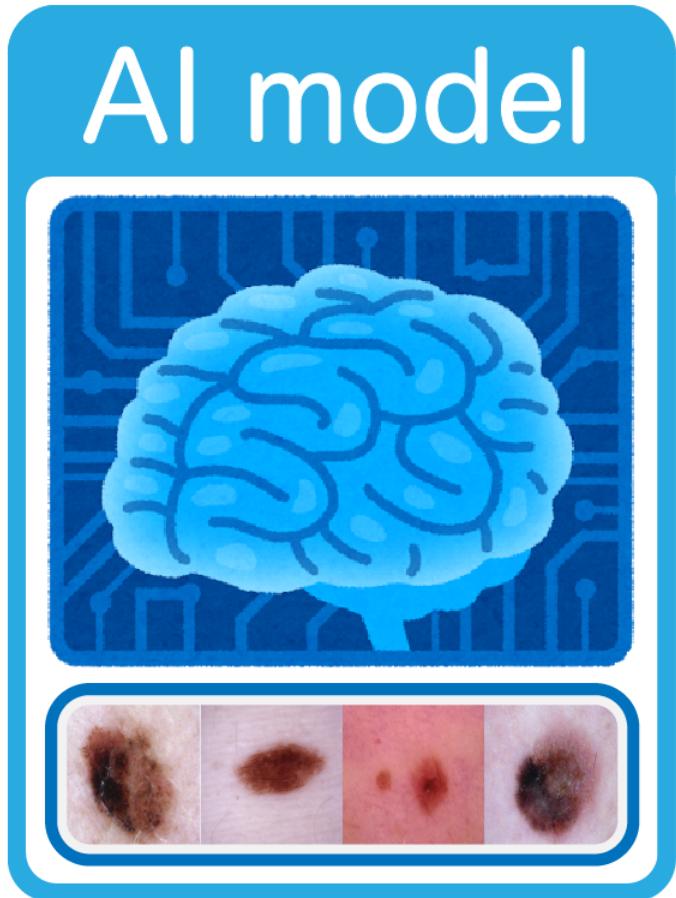
2<sub>17</sub>



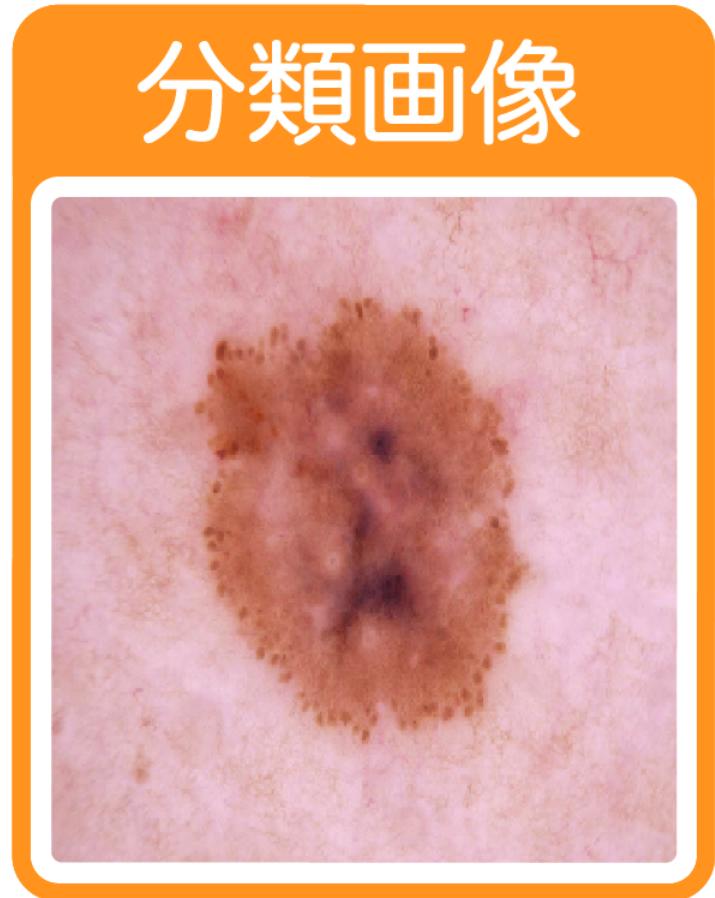
ほくろ → 皮膚がん

そのAIは、信頼できる？

# 敵対的擾動をつくるには？



&  
↓

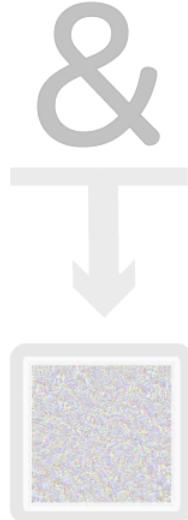
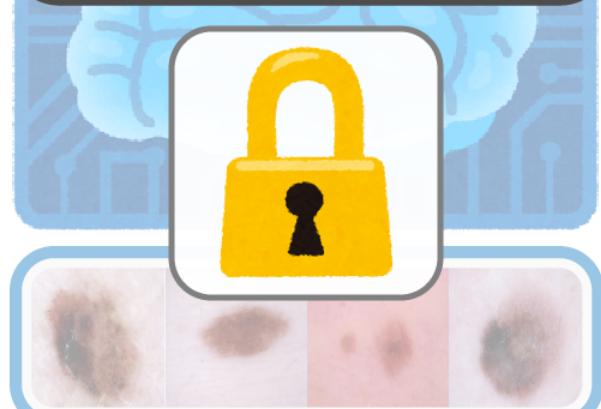


必要なのは、**モデルと分類画像**

# 敵対的擾動は、難しい。

AI model

入手困難



分類画像

入手困難

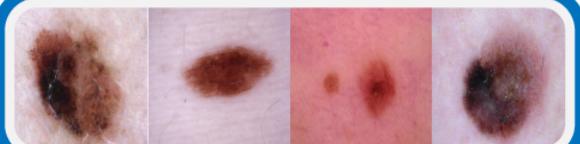
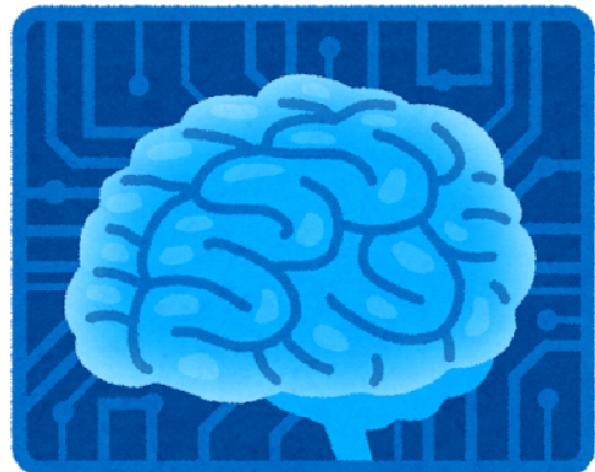


AI は信頼できそう？

# ところがどっこい

5<sub>17</sub>

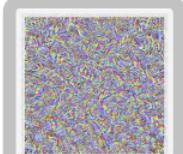
AI model



一般公開画像



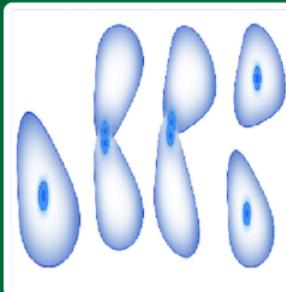
&  
↓



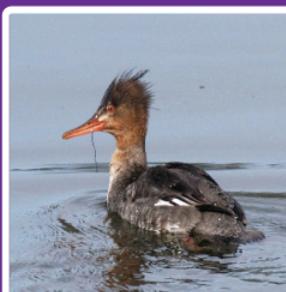
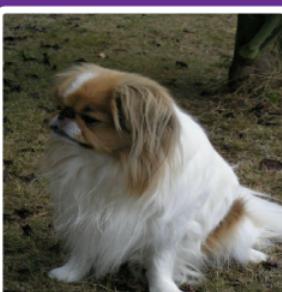
**分類画像は本当に必要？**

# 例えば、こんな画像

- Open Images Dataset V5



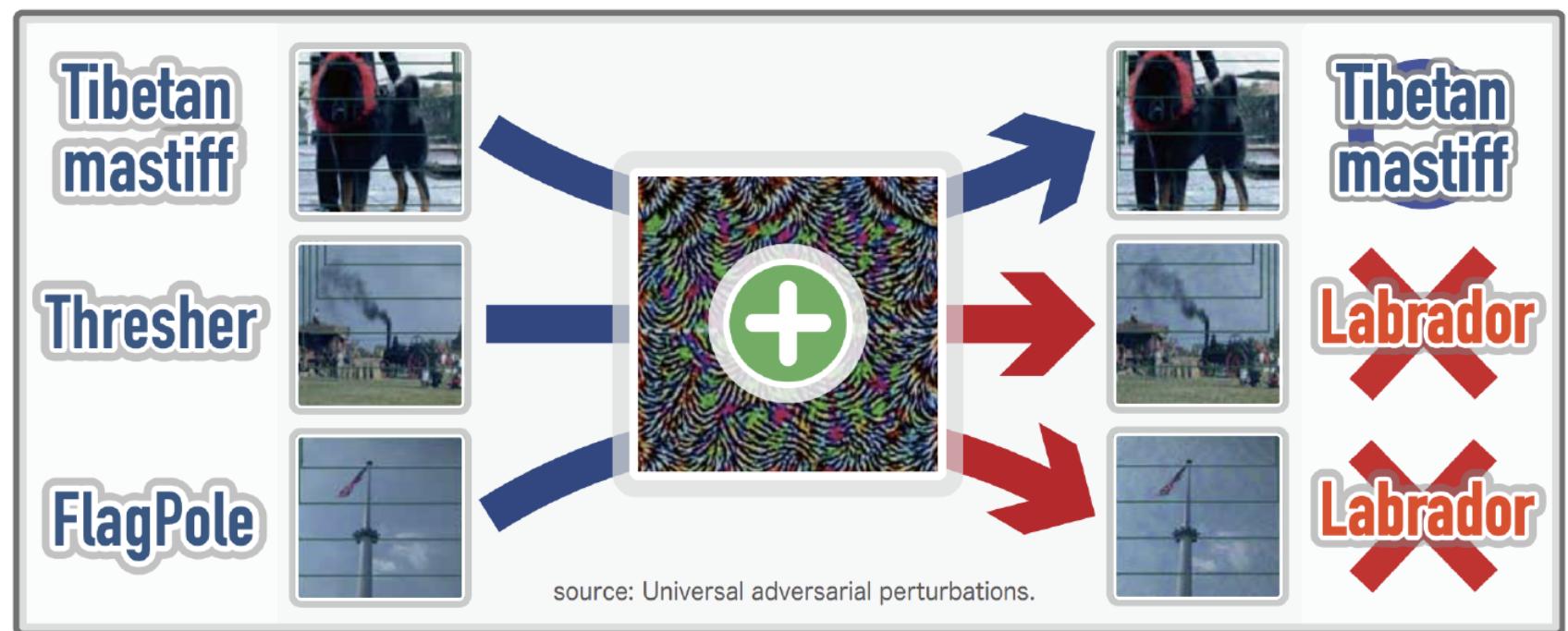
- ImageNet (ILSVRC2012)



誰でも、簡単に入手できる。

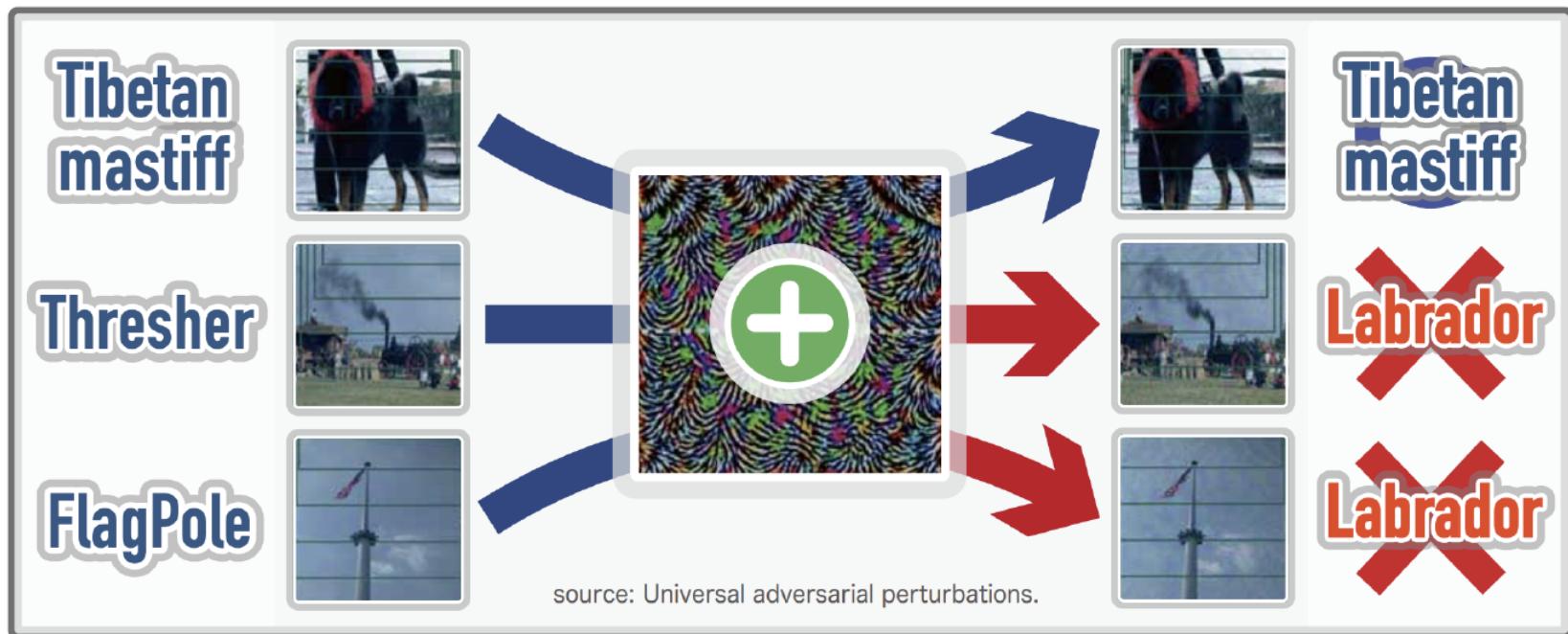
# ところで、普遍的撮動とは？

7<sub>17</sub>



1枚で画像をまとめて間違えさせる、  
特別な敵対的撮動のこと。

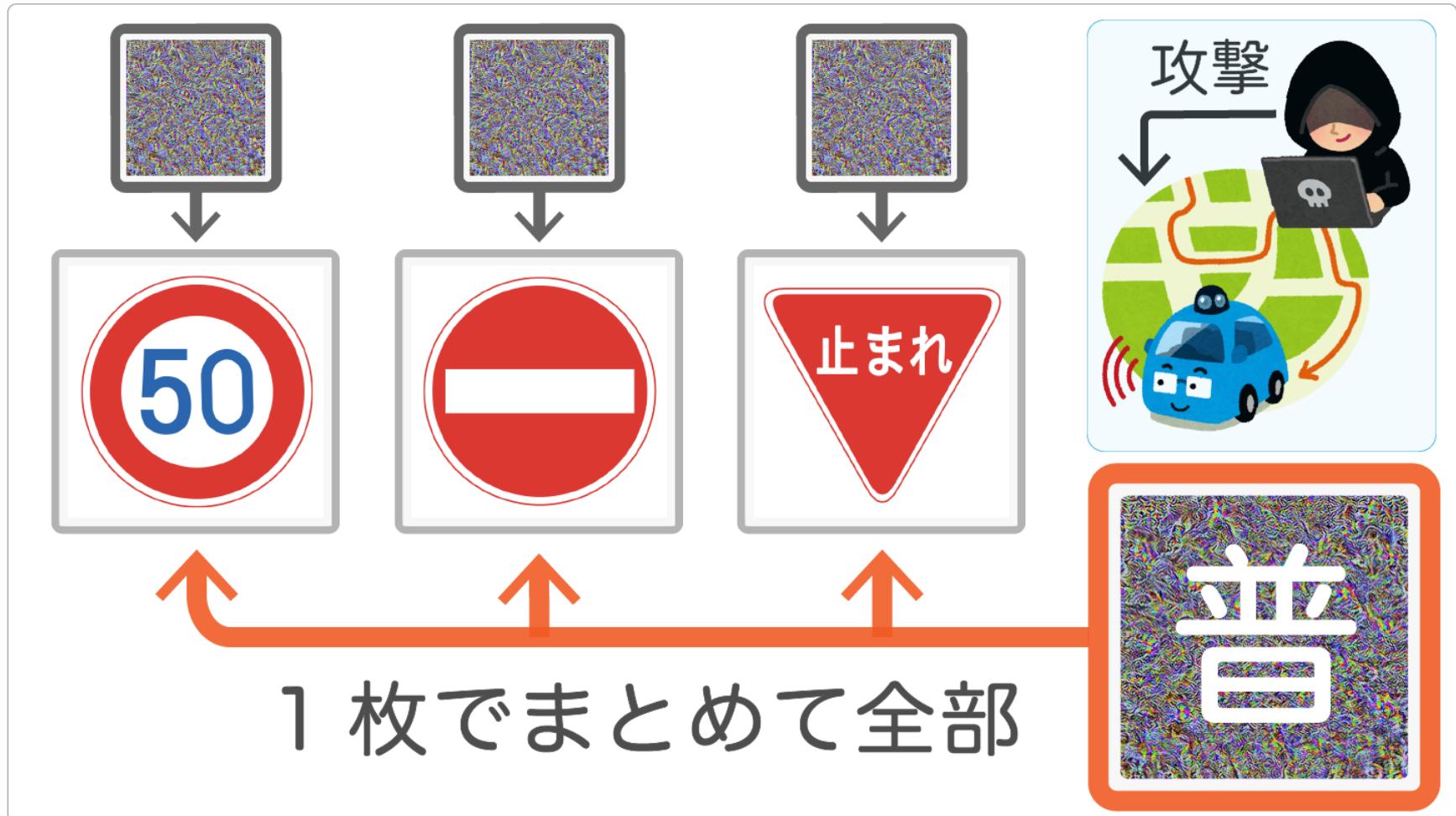
# ついでに、誤認識率とは？



誤認識率 = 2/3

間違えた画像の割合のこと。

# なぜ普遍的摂動？

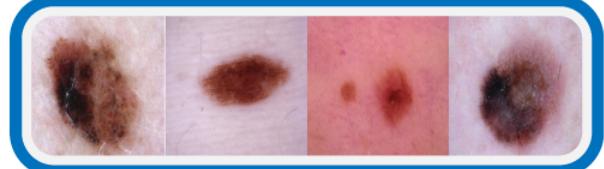
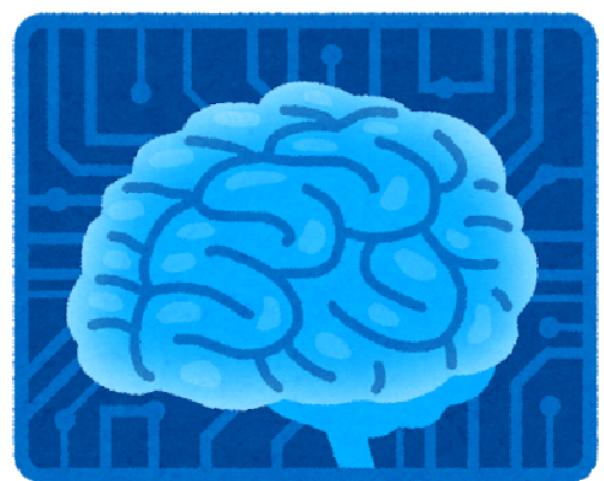


1つのAIを、1つの摂動で。

# 普遍的撮動をつくるには？

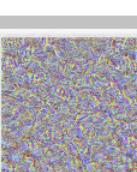
10<sub>17</sub>

AI model

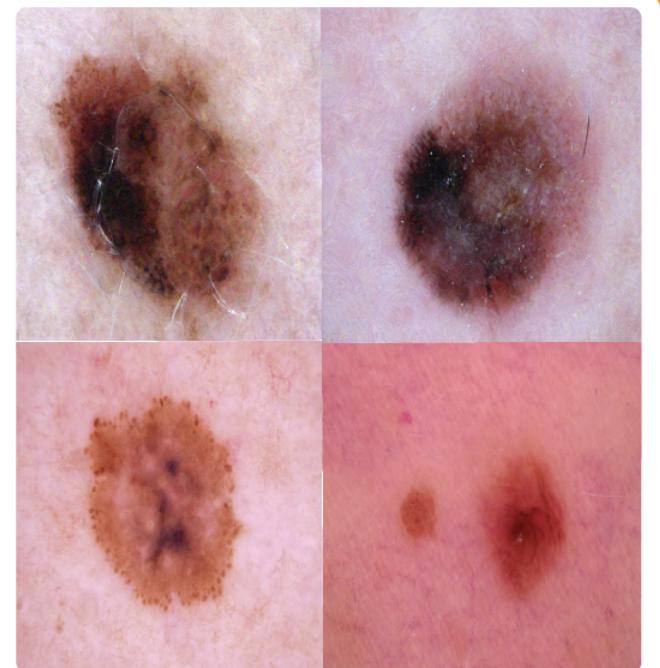


従来手法

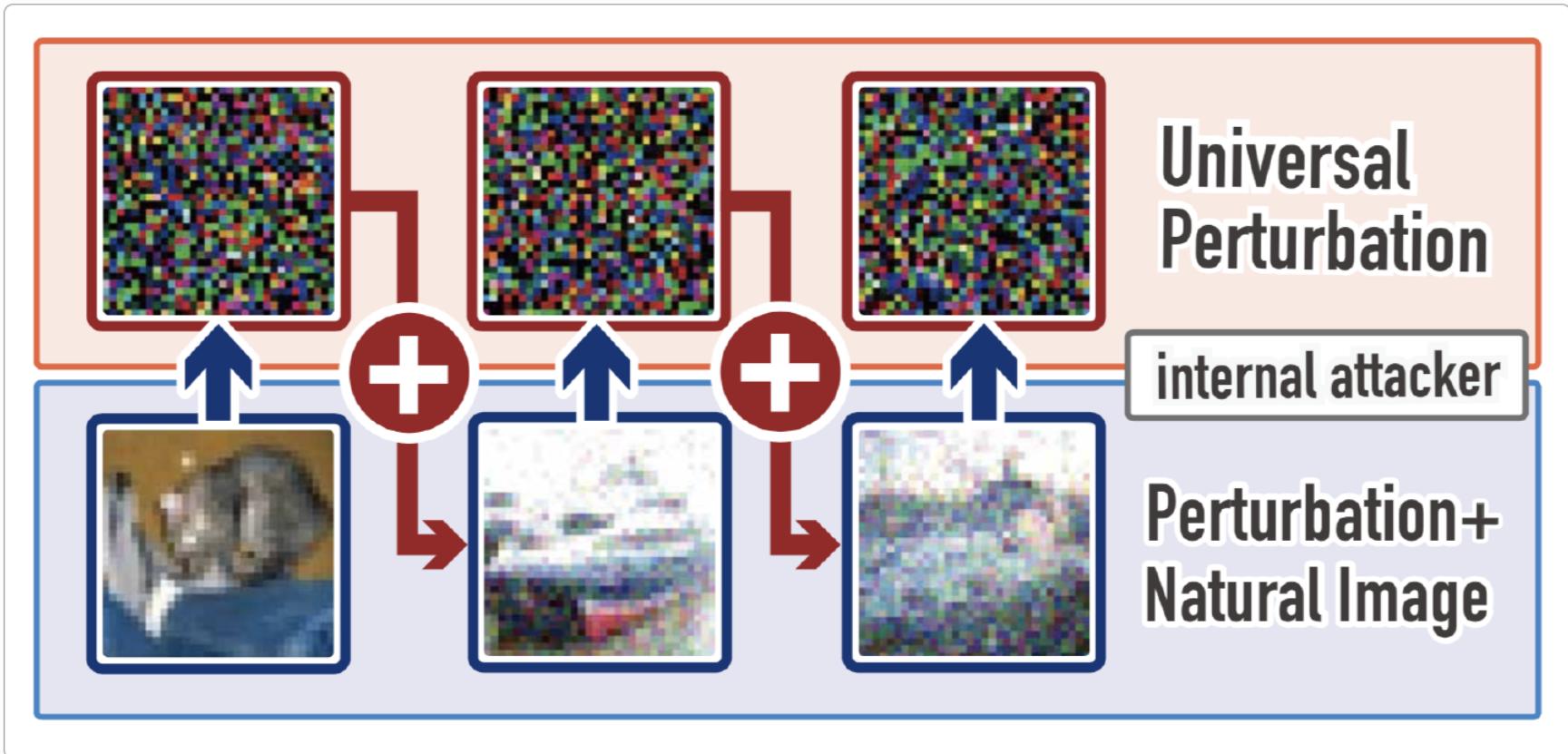
&



分類画像セット



モデルと分類画像の"セット"が必要。



普遍的摂動の作りかたは、  
敵対的摂動の生成の繰り返し。

# また、転移学習 AIについて

12<sub>17</sub>

- ・少ない訓練データで、高性能なモデルが得られるため、医療分野で頻繁に用いられる。
- ・ImageNet を事前学習した、InceptionV3, VGG16, ResNet50 を、ISIC2018（分類画像）で再訓練した。

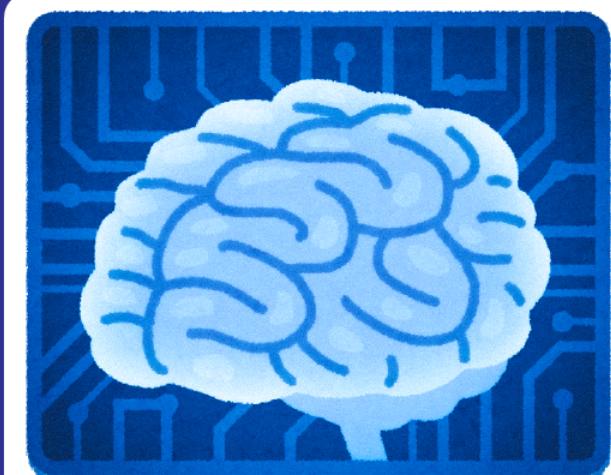
ディープニューラルネットワークによる皮膚癌の皮膚科医レベルの分類，Nature, 542:115-118, 2017.

検証には、転移学習 AIを採用。

# つまり、これを検証する。

13<sub>17</sub>

転移学習 AI



提案手法

&



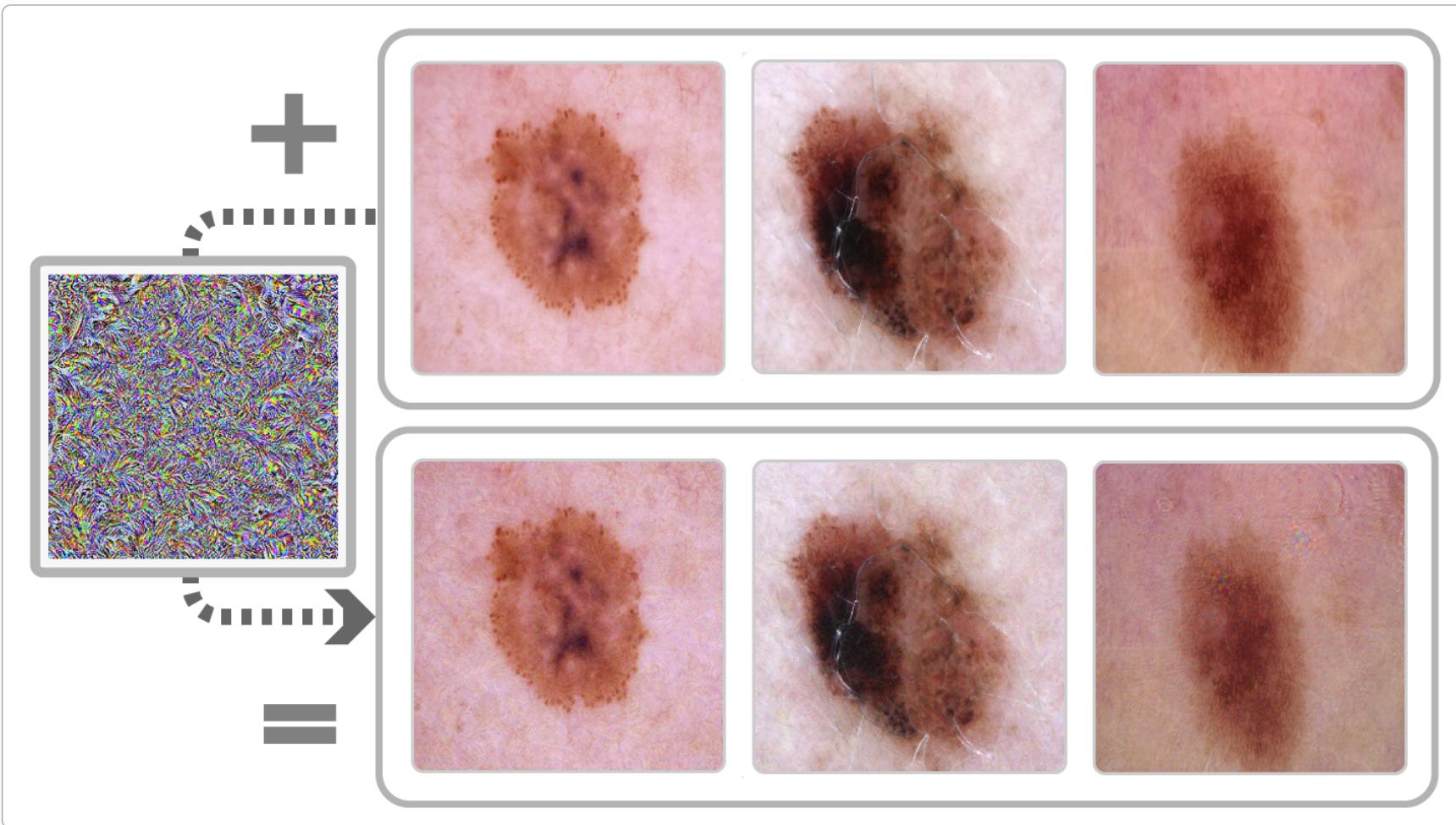
一般公開画像 セット



## 一般公開画像で、普遍的摂動をつくる。

# 普遍的撮動の生成は可能？

14<sub>17</sub>

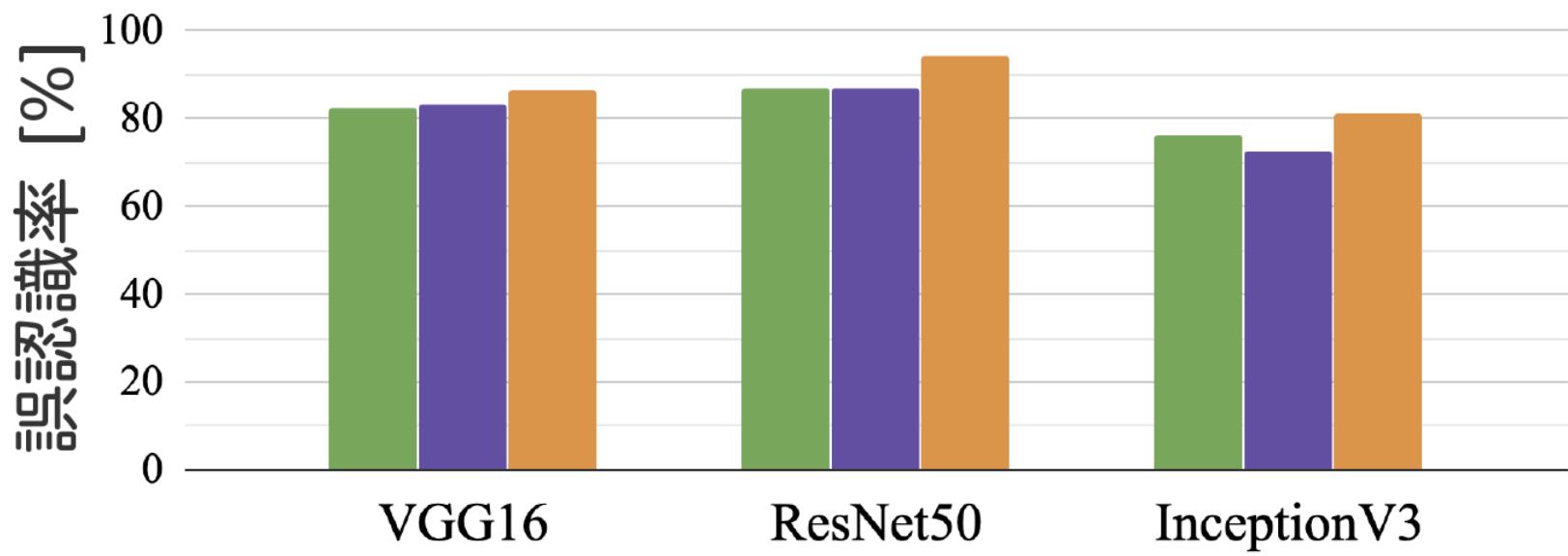


一般公開画像で、生成できた。



# その有効性は？

Dataset / Model	VGG16	ResNet50	InceptionV3
Open Images Dataset V5	82.5	86.7	76.3
ILSVRC2012	83.2	87.0	72.5
ISIC2018 (分類画像)	86.4	94.3	81.2

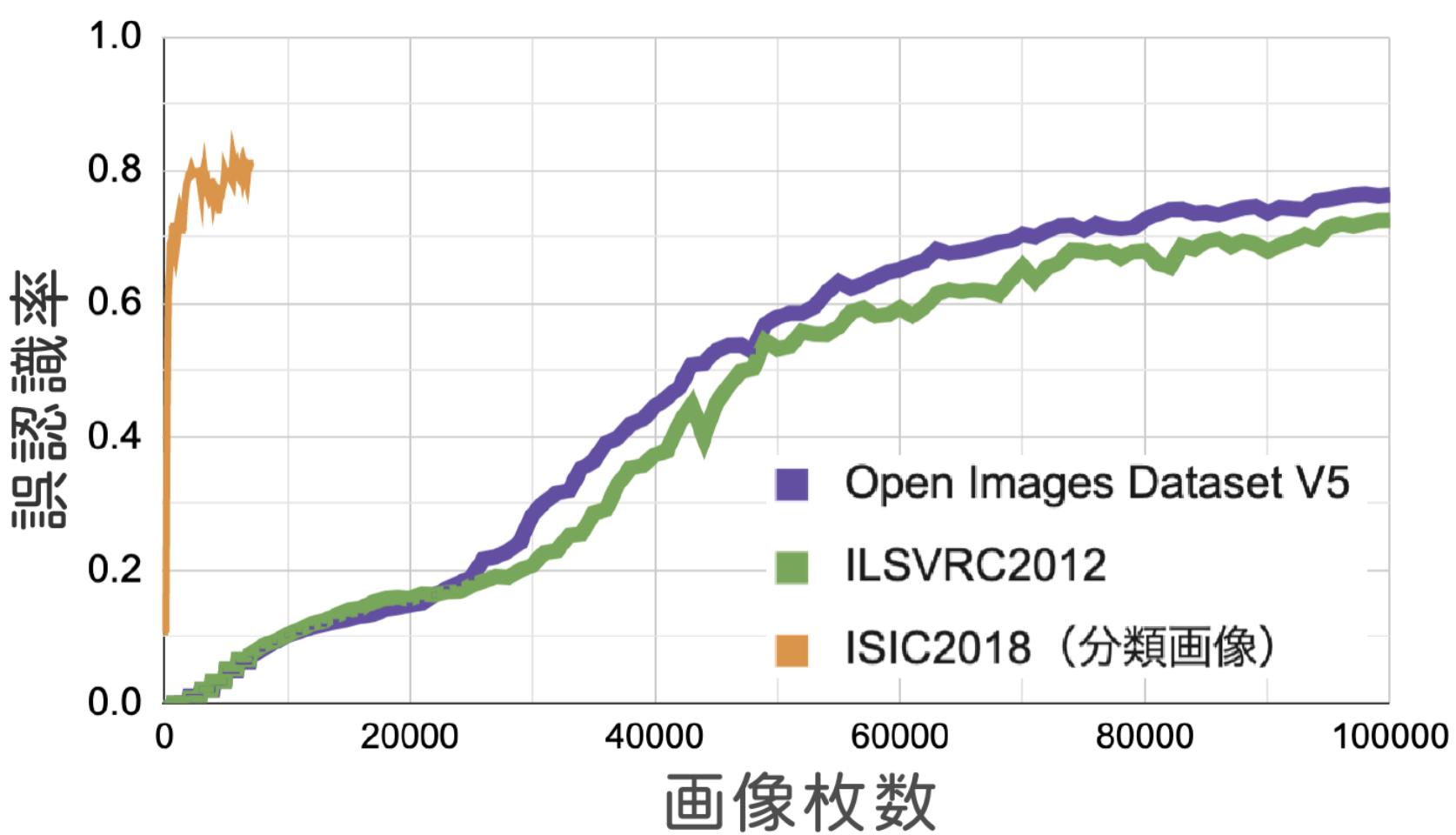


優れた誤認識率を達成した。



# その生成速度は？

16<sub>17</sub>



10倍以上の画像が必要。

課題

# まとめ

-- 今まで --

- AI は、敵対的撮動によって間違える。
- 敵対的撮動の生成には、  
AI モデルと分類画像が必要（入手困難）。

-- これから --

- 転移学習 AI モデルは、
- 一般公開画像から、敵対的撮動が生成可能。
- 転移学習の利用にはより慎重になるべきだ。