物理世界で発生しうる自然な摂動を利用した敵対的攻撃の研究

研究駆動コース　橋本 俊甫

1. 敵対的攻撃の存在

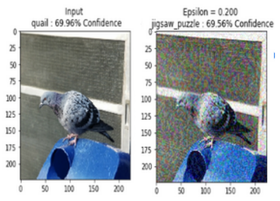
敵対的攻撃とは、機械学習モデルが誤った予測を行うように生成された入力 (敵対的サンプル) を使用することで、**意図的に誤認識を引き起こす**攻撃を指す。敵対的攻撃は、機械学習技術が重要な役割を果たすようになってきている現代社会においてますます深刻なセキュリティの脅威となっているため、機械学習技術を利用するシステムやサービスを開発する際には、その敵対的攻撃の存在を考慮して設計する必要がある。

2. これまでの研究

画像認識モデルを標的にした敵対的攻撃の研究は主に画像にランダムなノイズに見える摂動を付与する手法やカメラや被写体を細工するような手法である。誤認識をさせることを目標においた攻撃手法によって得られる入力画像は、**自然には発生しえない画像**や**物理世界に馴染めていない画像**になってしまう。

Fast Gradient Sign Method [1]

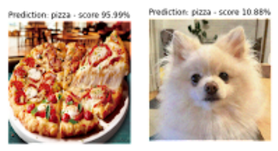
各ピクセルの勾配を調べてLossが大きくなる向きに応じて摂動を付与する
White-box型の攻撃手法



ランダムなノイズに
見える摂動

Boundary Attack [2]

最終的なモデルの出力結果のみを頼りに標的クラスを探索するように試行する
Black-box型の攻撃手法



うっすらと
ピザの面影が見える

Disappearance Attack [3]

物体検出器を標的に、
物理的な物体の位置を特定できなくしたり
存在しない物体を認識させたりできる
敵対的サンプルを生成

物理的な物体が検出器
から無視される

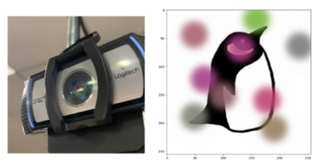
存在しないはずの物体の認識

3. アイデアの原石

物理世界以上の情報はデジタル世界へは持ち込めない。
物理世界の情報量を減らせば、その物体を撮影した画像は認識がしづらくなるはず。

Adversarial Camera Stickers [4]

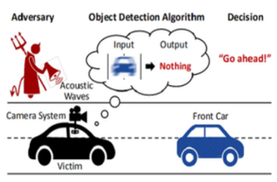
カメラ自体に細工をして、
カメラに映る映像自体をぼやけさせる
物理世界からの攻撃手法

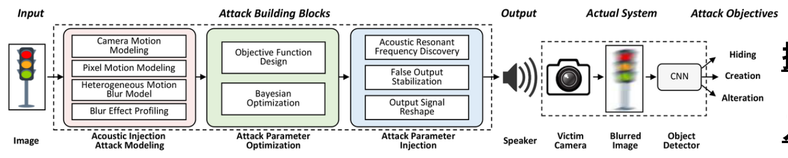


penguin → piggy bank
へ誤分類

Poltergeist Attack [5]

自動運転の文脈で、
慣性センサーに音響ノイズを加えることで
手ブレ補正機能を悪用できるようになる

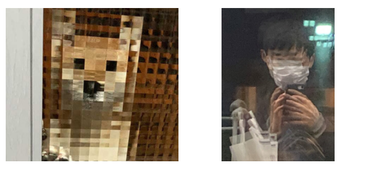




撮影する段階を攻撃すれば、後続の画像認識モデルの入力画像に細工せずに騙せる

物理世界を低解像度にする

目で見えない光・耳で聞こえない音を扱うような攻撃は確かにカッコイイがもっとカジュアルに攻撃できないか？



8 bit 犬 [6]　電車の窓に映る自分

物理世界に馴染む自然な敵対的攻撃手法は意外にシンプルで実現のヒントは日常に潜んでいるのかもしれない。

4. 物理世界に馴染む自然な摂動

脳が"バグる"マスク

画像処理で黒く塗りつぶされたかのように錯覚させることができれば、マスクで覆われた部分の情報が失われて認識精度は落ちるのだろうか？

真・黒色無双 [7]
（可視光の99.4%
を吸収する塗料）

マスクをしても
顔認識可能なモデル
→ ArcFace [8]
(Backbone: IR-152
Training Data:
MS-Celeb-1M)
→ 認識精度は特別落ちない

T-SNEで次元削減して
可視化した様子

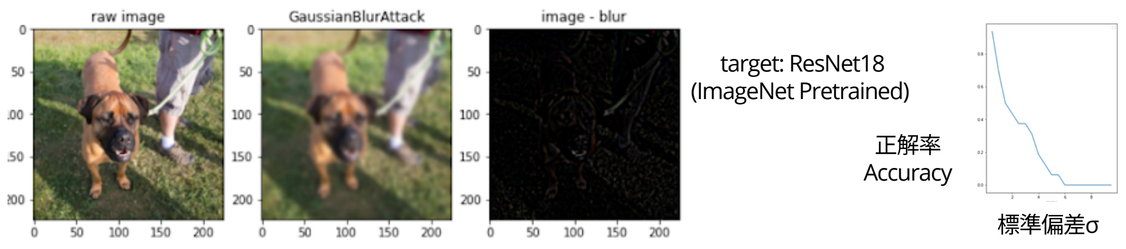
クラス内距離 ＜クラス間距離
になるよう配置されている

マスクをすると
顔認識不可能なモデル
→ dlibベースの
顔認識ライブラリ
Face Recognition [9]
→ 白いマスク着用時と同様に
顔の検出ができない

ブレを引き起こして敵対的サンプルと気づかせない

撮影時のブレ (Motion Blur) はカメラを手持ちで使っている場合や被写体が動いている場合など、さまざまな理由で**自然に生じることがあるため、敵対的サンプルであることを人間の目をも欺くことができる可能性がある。**

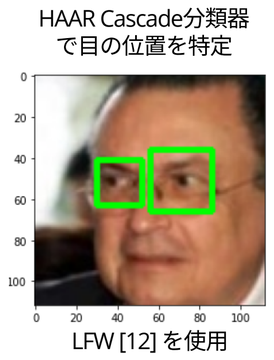
Gaussian Blur Attack [10] で画像の微細な特徴をぼかすことによって、誤認識を引き起こすことができる。標準偏差σの値を大きくするほど強いノイズになって認識精度は落ちるが、**小さい値でも認識精度に影響を与えられる場合がある。**



顔画像の目に注目することで探索範囲を限定する。
ブレの強さ・方向を最適化して最適なブレを探索しつつ、
元画像と敵対的サンプルの2つの類似性を測るSSIM [11] を最大化

$$\begin{aligned}&\text{adversarial\_loss}(\mathrm{x},\ \mathrm{x}_{adv}, blur\_direction, blur\_strength, w_m, w_s)\\&= w_m \cdot motion\_blur\_loss(x, x_{blur}, blur\_direction, blur\_strength)\\&+ w_s \cdot ssim\_loss(x, x_{adv})\end{aligned}$$

$$\begin{aligned}&\text{motion\_blur\_loss}(\mathrm{x},\ \mathrm{x}_{blur}, blur\_direction, blur\_strength)\\&= \frac{1}{HWD} * \sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} \left|\left|\frac{1}{K} * \sum_{k=1}^{K} x_{h,w,c,k} - x_{blurh,w,c}\right|\right|_2^2\\&\text{ssim\_loss}(\mathrm{x},\ \mathrm{x}_{adv}) = 1 - SSIM(x, x_{adv})\end{aligned}$$



デジタル世界における敵対的攻撃は、これまでにもさまざまな手法が提案されており、最近では物理世界からも攻撃可能であるかが重要になっている。物理空間での攻撃はデジタル空間での攻撃に比べると難しい場合があり、目をブレさせるアイデアはやや現実的でないとしても、**理論的に攻撃可能であることを示すことは、実際の攻撃が行われる前に対策や防御策を考えるための重要なステップとなる。**

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
[2] Brendel, Wieland, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." arXiv preprint arXiv:1712.04248 (2017).
[3] Eykholt, Kevin, et al. "Physical adversarial examples for object detectors." arXiv preprint arXiv:1807.07769 1.3 (2018): 4.
[4] Li, Juncheng, Frank Schmidt, and Zico Kolter. "Adversarial camera stickers: A physical camera-based attack on deep learning systems." International Conference on Machine Learning. PMLR, 2019.
[5] Ji, Xiaoyu, et al. "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision." 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021.
[6] https://twitter.com/i/status/1435743602722820098, [7] https://www.amazon.co.jp/dp/B08JYK85HP, [9] https://github.com/ageitgey/face_recognition
[8] Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[10] Zhao, Chenchen, and Hao Li. "Blurring Fools the Network--Adversarial Attacks by Feature Peak Suppression and Gaussian Blurring." arXiv preprint arXiv:2012.11442 (2020).
[11] Zhao, Hang, et al. "Loss functions for image restoration with neural networks." IEEE Transactions on computational imaging 3.1 (2016): 47-57., [12] http://vis-www.cs.umass.edu/lfw/