


ДОДАТОК А
Слайди презентації



Дослідження методів змагальних атак на нейронні мережі

Виконала: ст. гр. ПЗм-17-1 Сізон Ю.О.

Керівник роботи: доц. каф. ПІ, к.т.н. Каук В.І.

Рисунок 1 – Титульний слайд



Змагальні атаки

У 2014 році група дослідників штучного інтелекту із Google винайшли нову вразливість нейронних мереж, що назвали змагальною атакою.

Змагальні атаки – техніки, спрямованих на те, щоб обманути модель через подання зловмисних даних, тобто **змагальних прикладів**.

Змагальний приклад – зразок з малими, навмисними ускладненням, зашумленням ознак, які змушують модель машинного навчання робити помилкове передбачення. Змагальні приклади є даними, що подаються на вхід до моделі та мають за мету обманути її.

Рисунок 2 – Змагальні атаки

Змагальні атаки на системи розпізнавання зображень

Змагальні приклади для зображень – це зображення з навмисно зміненими пікселями з метою обманути модель під час застосування. Приклади наочно демонструють, наскільки легко глибокі нейронні мережі для розпізнавання об'єктів можуть бути обмануто зображеннями, які здаються адекватними для людини. Змагальні приклади подібні до оптичних ілюзій, але для комп'ютерів.

До методів змагальних атак відносяться такі як метод 1-піксельної атаки, метод змагального патча, метод чорної скриньки та інші, в тому числі метод стійких змагальних прикладів, що був застосований у роботі і буде розглянутий далі.

Рисунок 3 – Змагальні атаки на системи розпізнавання зображень

Архітектура Xception

1. Згорнути вихідний тензор 1×1 згорткою, отримавши тензор $M * M * C2$. Ця операція називається *pointwise convolution*.

2. Згорнути кожен канал окремо 3×3 сверткою (при цьому розмірність не зміниться, так згортаються не всі канали разом, як в звичайному згорточному шарі). Ця операція називається *depthwise spatial convolution* (глибинна просторова згортка).

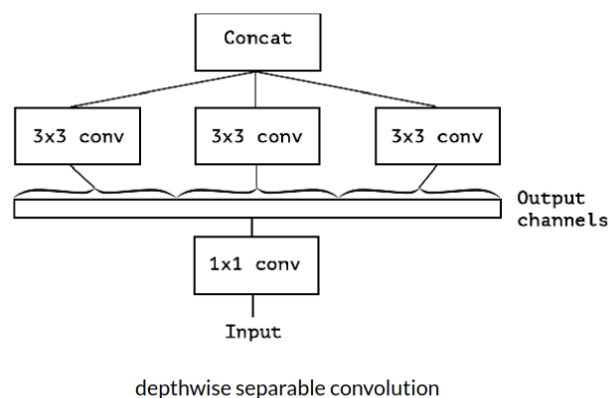


Рисунок 4 – Архітектура

Результати тренування



0.9615, cat



0.9221, dog

Рисунок 5 – Результати тренування

Створення змагального прикладу

Маючи зображення x , нейронна мережа виводить розподіл ймовірностей над мітками, $P(y/x)$.

Коли розробляється конкурсний вхід, потрібно знайти \hat{x} , де $\log P(\hat{y}|\hat{x})$ максимізовано для цільової мітки \hat{y} . Можна переконатися, що \hat{x} не виглядає занадто відмінним від початкового x , із обмеженням ℓ_∞ , радіусом ϵ , щоб $\|x - \hat{x}\|_\infty \leq \epsilon$.

Спочатку проводимо ініціалізацію змагального прикладу як $\hat{x} \leftarrow x$. Потім повторюємо приведені нижче кроки до збіжності:

$$\hat{x} \leftarrow \hat{x} + \alpha \cdot \nabla \log P(\hat{y}|\hat{x})$$

$$\hat{x} \leftarrow \text{clip}(\hat{x}, x - \epsilon, x + \epsilon)$$

Далі потрібно описати крок градієнтного спуску, щоб максимізувати логарифмічну ймовірність цільового класу. Необхідно описати крок проєкції, щоб змагальний приклад був візуально близьким до вихідного зображення

Рисунок 6 – Створення змагального прикладу

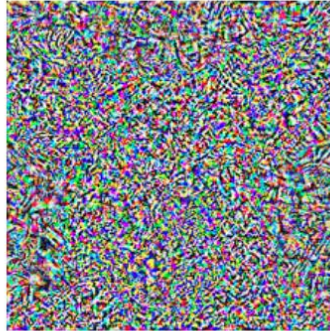
Результати атаки

```
the step is 10, the loss is 3.95831
the step is 20, the loss is 0.571342
the step is 30, the loss is 0.0315487
the step is 40, the loss is 0.0214902
the step is 50, the loss is 0.0173401
the step is 60, the loss is 0.0143289
the step is 70, the loss is 0.0121266
the step is 80, the loss is 0.0111390
the step is 90, the loss is 0.00954381
the step is 100, the loss is 0.0089127021
```



0.9615, cat

+



=



0.9008, dog

Рисунок 7 – Результати атаки

Покращення методу атаки



0.8502, cat

Рисунок 8 – Покращення методу атаки

Покращення методу атаки

З урахуванням деякого ротаційного перетворення T можна максимізувати $E_{t \sim T} \log P(\hat{y} / t(\hat{x}))$, враховуючи $\|x - \hat{x}\|_{\infty} \leq \epsilon$. Цю задачу оптимізації можна вирішити за допомогою проектного градієнтного спуску, зазначивши, що

$$\nabla E_{t \sim T} \log P(\hat{y} / t(\hat{x})) E_{t \sim T}$$

є:

$$\nabla \log P(\hat{y} / t(\hat{x}))$$

і наближається до зразків на кожному кроці градієнтного спуску.

Рисунок 9 – Покращення методу атаки

Результати

```
the step is 50, the loss is 0.0803217
the step is 100, the loss is 0.0221388
the step is 150, the loss is 0.00732416
the step is 200, the loss is 0.00391829
the step is 250, the loss is 0.00626519
the step is 300, the loss is 0.00234199
```

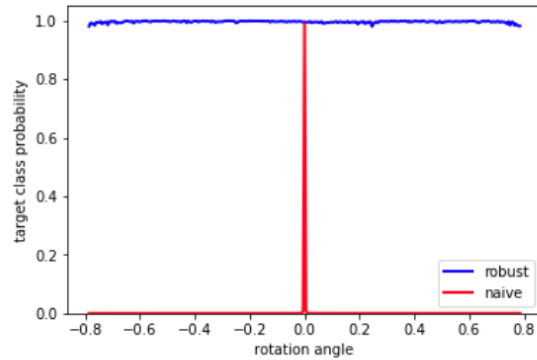


0.9514, dog

Рисунок 10 – Результати

Результати

$P(\hat{y} \neq x)$ для $\theta \in [-\pi/4, \pi/4]$



успішність атаки, залежно від куту ротації змагального приклада

Рисунок 11 – Результати

Користь роботи, висновки

Результати розробленого удосконаленого методу змагальної атаки вказують на поточні вразливості нейронних мереж та потенціал для зловмисного використання, тому результати роботи можуть послужити у проектуванні більш надійних нейронних мереж, що є безпечнішими для впровадження у системи, що працюють із розпізнаванням зображень.

Рисунок 12 – Висновки