ДОДАТОК А

Слайди презентації

Рисунок 1 – Титульний слайд

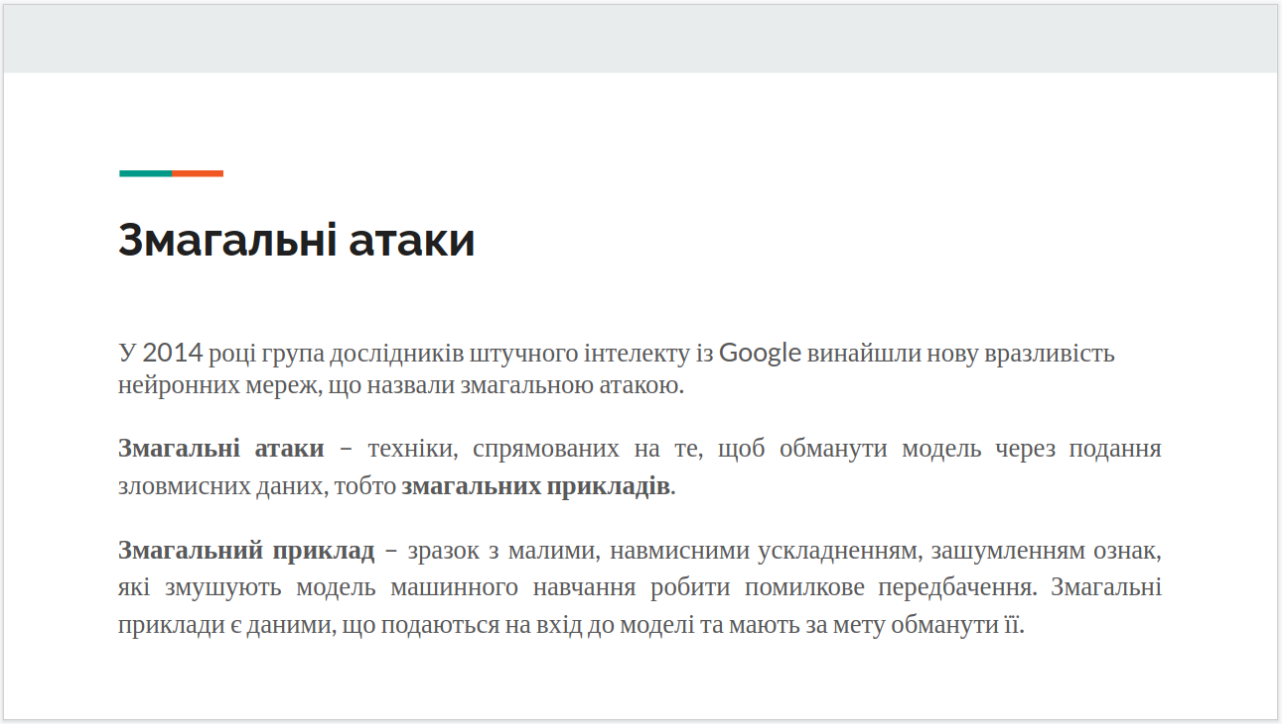


Рисунок 2 – Змагальні атаки

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.

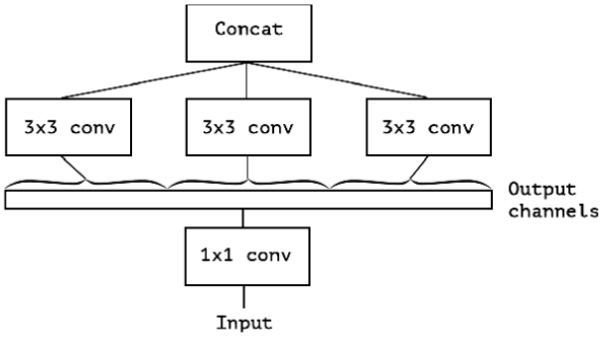
До методів змагальних атак відносяться такі як метод 1-піксельної атаки, метод змагального патча, метод чорної скриньки та інші, в тому числі метод стійких змагальних прикладів, що був застосований у роботі і буде розглянутий далі.

Рисунок 3 – Змагальні атаки на системи розпізнавання зображень

Архітектура Xception

1. Згорнути вихідний тензор 1x1 згорткою, отримавши тензор M * M * C2. Ця операція називається *pointwise convolution*.

2. Згорнути кожен канал окремо 3x3 сверткой (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).



depthwise separable convolution

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 $logP(\hat{y}|\hat{x})$ 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:

$$\hat{x} \leftarrow \hat{x} + \alpha \cdot \nabla logP(\hat{y}|\hat{x})$$
$$\hat{x} \leftarrow clip(\hat{x}, x - \epsilon, x + \epsilon)$$

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

Рисунок 6 – Створення змагального прикладу

Рисунок 7 – Результати атаки

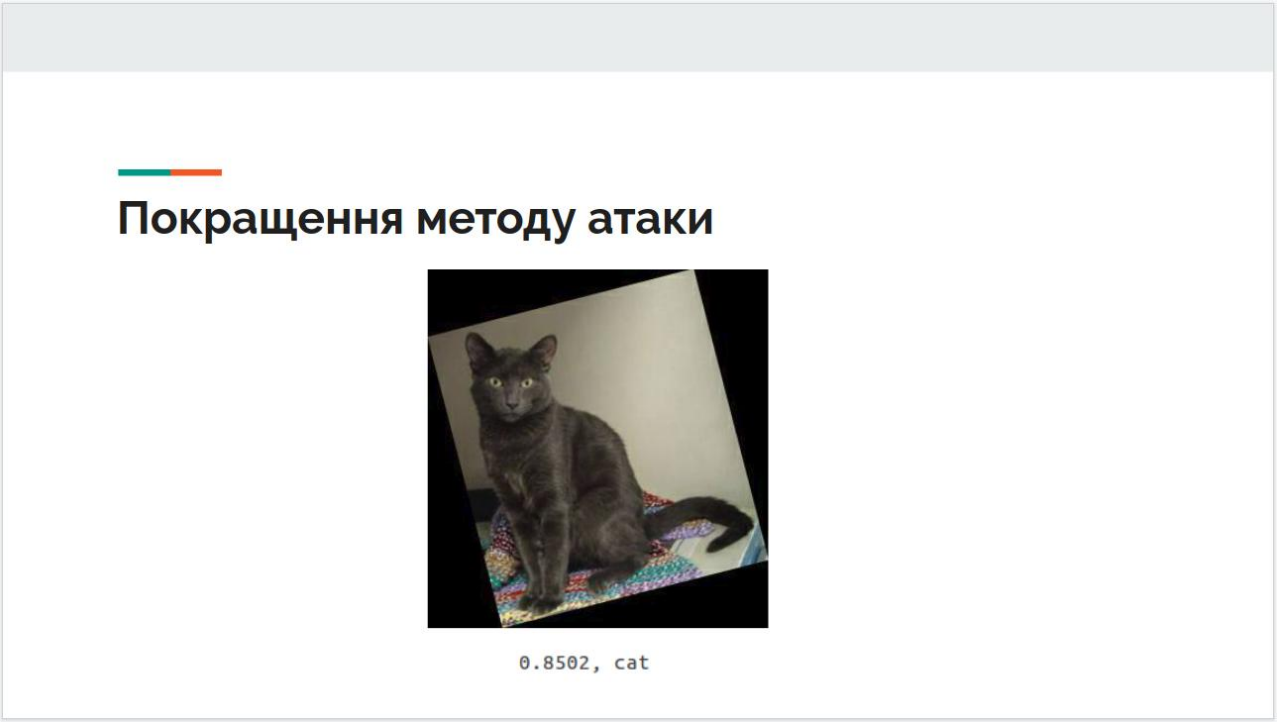


Рисунок 8 – Покращення методу атаки

Покращення методу атаки

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

$$\nabla\, \mathbb{E}_{t\sim T} logP\ (\hat{y}\ |\ (t\ (\hat{x}))\ \mathbb{E}_{t\sim T}$$

є:

$$\nabla\, logP\ (\hat{y}\ |\ t\ (\hat{x}))$$

і наближається до зразків на кожному кроці градієнтного спуску.

Рисунок 9 – Покращення методу атаки

Результати

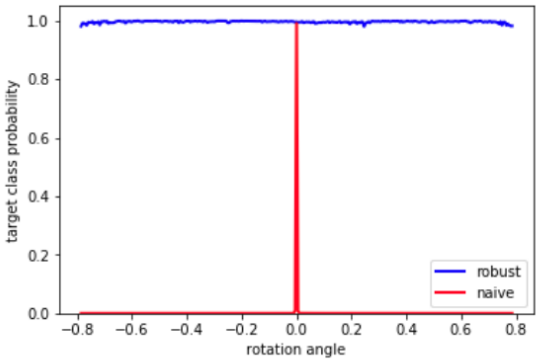
```
the step is 50, the loss is 0.0803217
the step is 100, the loss is 0.0221388
the step is 150, the loss is 0.00732416
the step is 200, the loss is 0.00391829
the step is 250, the loss is 0.00626519
the step is 300, the loss is 0.00234199
```

0.9514, dog

Рисунок 10 – Результати

Результати

$P(\hat{y} \mid x)$ 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