Перенесення адаптивної анонімізації в задачі класифікації зображень

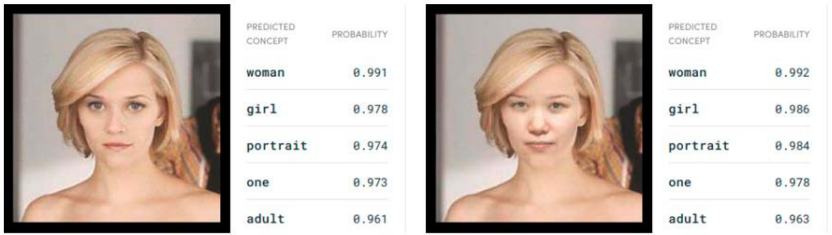
Актуальність роботи

Алгоритми генеративного штучного інтелекту в наш час перебувають в апогеї свого розвитку і в свою чергу вимагають великої кількості зображень для коректної роботи.

Негативний ефект анонімізації зображень на роботу моделі

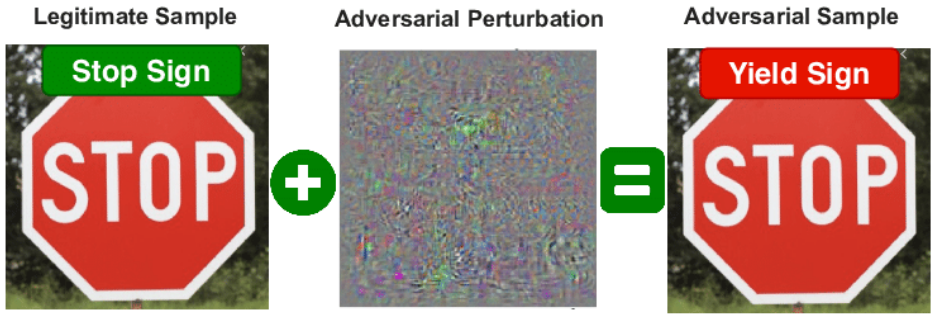
Однак виконання цієї норми шляхом використання наївних методів анонімізації часто призводить до значного погіршення результатів роботи класифікаційних нейронних мереж.

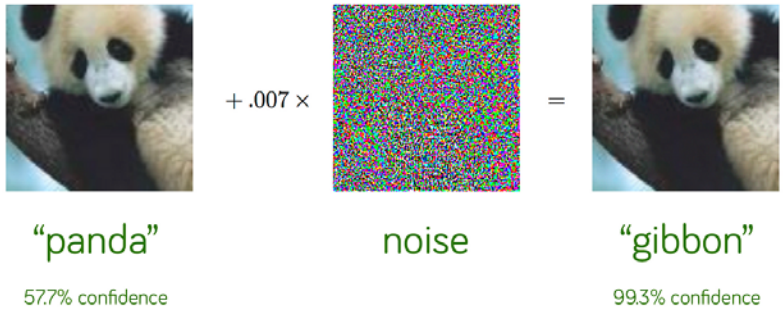
Методи анонімізації зображень



- Використання Генеративно Змагальних Мереж для анонімізації облич.
- Використання підходу “Зворотня Супер Роздільна Здатність” (ISR).Така роздільна здатність, безумовно, дозволяє зберегти приватність даних.
- Градієнтний метод ітеративної модифікації зображення.

Змагальні атаки





Явище переносимості змагальних прикладів

Prediction based on AE with original label “cat” 

	(1)	(2)	(3)
F1: Source model	✗: “dog”		
F2: Target model	✓: “cat”	✗: “frog”	✗: “dog”

Definitions of adversarial transferability

Non-targeted transferability	Non-transferable	Transferable	
Targeted transferability (target class: “dog”)	Non-transferable		Transferable
Class-aware transferabiltiy	Unfooled	Different mistake	Same mistake

Переносимість адаптивної анонімізації

Так як задача адаптивної анонімізації по суті є протилежною до задачі створення змагальних атак.
Ми вирішили перевірити базовий градієнтний алгоритм на явище переносимості.

Назва моделі	Точність класифікації
MobileNetV2_100 (базова модель)	100%
ResNet52	60.6%
ResNet101	69.1%
ResNet152	71.9%
VGG13	30%
VGG16	29.4%
VGG19	30.8%
MobileNetV2_050	32.9%
MobileNetV2_140	41.8%
MobileNetV3_large_100	19.8%

Переносимість адаптивної анонімізації

Наступним кроком було вирішено, спробувати покращити базовий алгоритм, розширивши його застосування на декілька моделей водночас.

MobileNetV2_100 (базова модель)	100%
MobileNetV2_140	100%
ResNet101	100%
ResNet152	100%
VGG16	100%
VGG19	100%
Xception61	100%

Дякую за увагу