UNIVERSITÉ DE MONTPELLIER

Compte rendu 2 - Etat de l'art

BELDJILALI Maxime, CHATEAUNEUF Arthur
13 October 2024

Introduction

Pertinence de l'obscuration

Depuis plus d'un siècle, les chercheurs en mathématique, en informatique et parfois même en biologie ont imaginé des systèmes de recréation de neurones artificiels. Malgré des années de travail longues et ardues, cela ne fait qu'une à deux décennies que ces systèmes théoriques ont pu être mis en place, testés voire utilisés par le grand public.

La présence toujours plus accrue de ces intelligences artificielles, ainsi que la croissance de leur puissance et de leurs capacités mènent régulièrement à de nouveaux problèmes contemporains. Que cela concerne la sécurité de nos infrastructures, la protection de nos vies privées ou même la maintenance de notre patrimoine artistique, il est plus que jamais important de fonder notre avenir sur la recherche autour de ces questions.

Le sujet que nous souhaitons aborder est celui de l'obscuration d'image, qui consiste à cacher une information, dans notre cas visuelle, à un ou plusieurs observateurs. Cela peut servir à étudier efficacement les meilleurs moyens d'anonymiser les contenus multimédias

(comme des témoignages, des reportages etc…), empêcher le trackage abusif de personnes sur le net ou même protéger des œuvres de vols en masse.

Problèmes de recherche

La recherche autour de l'obscuration de média et des réseaux de neurones convolutifs (CNN) possède de nombreuses complications. Premièrement, la complexité des CNN ainsi que leur évolution constante rend difficile l'évaluation de la robustesse et peut même entraîner des méthodes d'obscuration trop lourdes en calculs ou en fidélité. Il faut, par ailleurs, prendre en compte tous les aspects de la manipulation de données multimédias, c'est-à-dire être conscient de la nature des perceptions de chaque observateur (humains, algorithmes statiques, CNN etc…) et des limitations de l'utilisation souhaitée (taille pour partage sur le net, vitesse d'utilisation, prérequis externes etc…)

Ce domaine est ainsi en pleine expansion et doit répondre à des problèmes d'actualité tout en maintenant un équilibre important entre plusieurs composantes essentielles.

Exemples d'approches

Obscurations naïves

Le premier papier que nous allons utiliser dans cet état de l'art est "Defeating Image Obfuscation with Deep Learning" par Richard McPherson, Reza Shokri et Vitaly Shmatikov en Septembre 2016. Ce document étudie les performances de méthodes modernes de reconnaissances d'images par CNN face aux cas naïfs d'obscuration tels que le floutage, la pixellisation ou la manipulation d'algorithmes de compressions (comme démontré avec les coefficients JPEG).

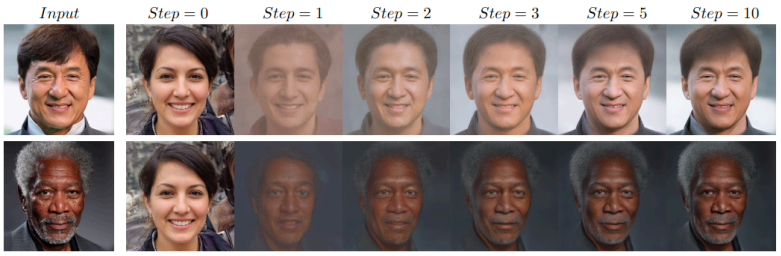
La conclusion de leurs résultats est que des réseaux de neurones sont capables de s'adapter à la dégradation de données pures, c'est-à-dire à une modification globale des données de l'image. Nous appelons ainsi ces obscurations naïves car elles ont pour but de retirer des informations de façon uniforme sans objectifs précis sur la différenciation des observateurs. Ces dégradations entraînent la majorité du temps une perte de fidélité pour tous les observateurs, de façon indiscriminée. Cependant, le papier montre que les CNN sortent souvent gagnants de ce genre de méthodes.

DeepBlur

Le second papier qui nous intéresse, "DeepBlur: A Simple and Effective Method for Natural Image Obfuscation" par Tao Li et Min Soo Choi prend une approche bien plus novatrice. Leur méthode consiste à effectuer un floutage dans l'espace tangent d'une image. Les données manipulées ne sont ainsi plus des signaux multimédias mais des concepts plus abstraits comme la description des caractéristiques d'un visage.

Leur algorithme itératif intègre directement des réseaux de neurones, que ce soit pour l'extraction des données tangentes ou la génération de l'image de sortie. Cette approche

permet de garder une forte reconnaissance visuelle pour un observateur humain tout en empêchant un CNN de reconnaître l'identité du visage.



Exemple donné de résultats d'obscuration de photo de célébrités à partir d'un visage moyen

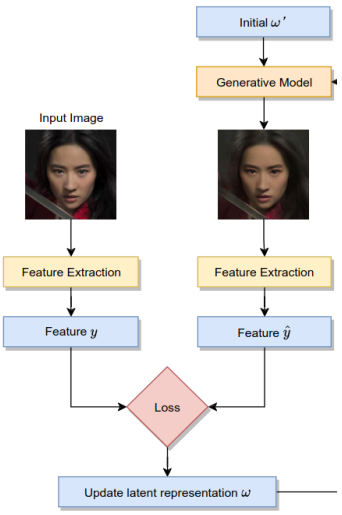


Figure 3. A general framework of latent representation sea

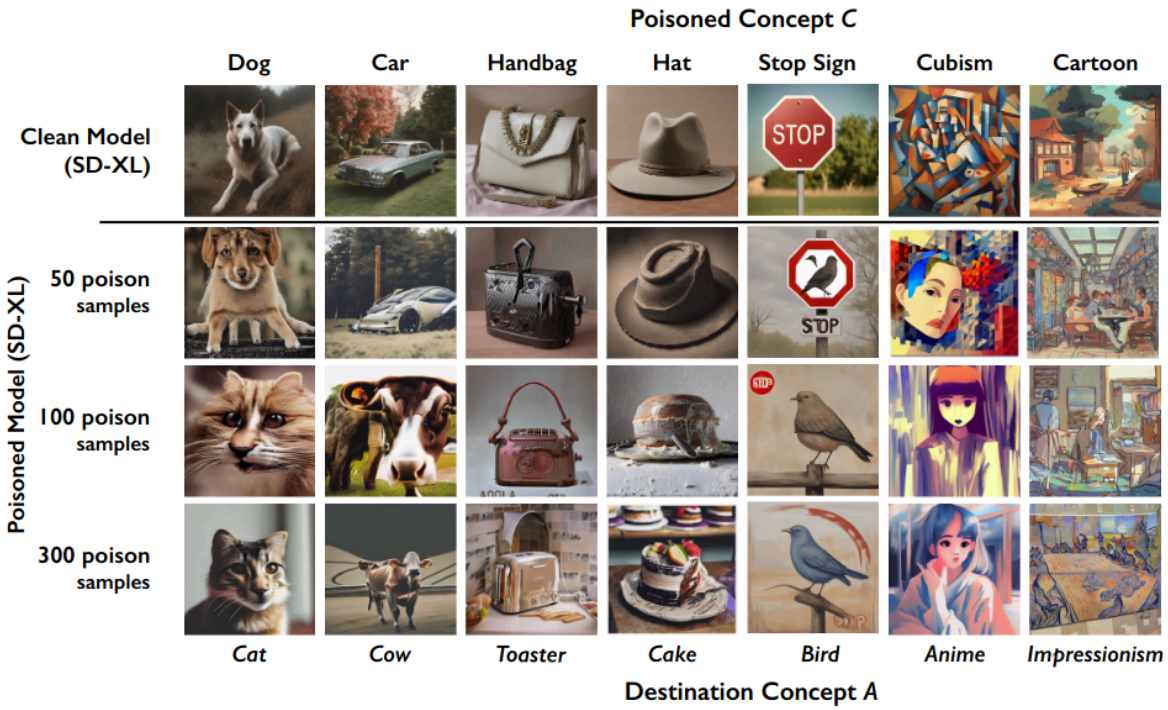
NightShade

Le dernier article qui nous intéresse est "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models" par Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng et Ben Y. Zhao.

Ce papier parle de techniques d'empoisonnement de réseau de neurones. Cela consiste à altérer les données d'entraînements ou de tests de CNN afin de leur induire des erreurs. Leur méthode, qui s'attaque directement aux IA génératives d'images, est d'une efficacité particulière. Elle se base sur un principe simple : Bien que les données d'entraînement et de test de ces IA soient énormes, contenant parfois des milliards d'images, la quantité énorme d'informations et de concepts tangents qu'elles doivent manipuler permet des attaques précises.

Les chercheurs parlent d'une approximation de seulement quelques milliers d'images qui peuvent servir pour constituer un seul concept dans le réseau.

Leur approche est la suivante : détecter la catégorisation d'une image et lui appliquer des modifications imperceptibles à l'œil nu en rapport avec une autre catégorie. Dans leur exemple, ils empoisonnent des images de chiens avec des données extraites d'images de chats. Cette attaque par confusion de concepts s'avère très efficace pour détruire de larges CNN avec un petit nombre de donnés empoisonnés.



Exemple de résultats donnés dans le papier

Cette approche prend ainsi le problème à l'envers, et cherche à empêcher des modèles compétents d'être créés, maintenus et améliorés. Cela peut être très pertinent lorsque l'on met en exergue la quantité toujours plus élevée de données que les CNN requiers.

Critiques & analyses

Fidélité visuelle

Comme nous l'avons expliqué, la fidélité visuelle des obscuration naïves devient très pauvre lorsque l'on essaie de viser un type précis d'observateur.

Les algorithmes basés tangents, quant à eux, donnent des résultats que nous considérons aléatoires au mieux. Bien que la reconnaissance visuelle de la caractéristique floutée précise soit bonne, des changements étranges dans les images sont inévitables. Nous avons, par exemple, les arrières plans perdant énormément de fidélité, les détails d'éclairages totalement perdus ou même des visages qui semblent s'approcher d'une moyenne (exemple : des rides qui disparaissent totalement, changeant l'âge apparent d'un visage). De façon générale, nous considérons que l'utilisation de réseaux de neurones générateurs obtiendra toujours des résultats à l'apparence étrangement lissée et moyennée.

Les meilleurs résultats de fidélité sont obtenus avec la méthode Nightshade. L'idée de ne pas naïvement détériorer une grande partie de l'information, et de ne pas avoir à recréer une réplique de l'image, mais d'au contraire gardé un maximum d'informations tout en visant les attributs utilisés par un type précis d'observateur entraîne des pertes visuelles totalement nulles.

Robustesse

Comme nous l'avons vu, la robustesse des obscurations naïves est limitée due à leur absence de précision sur les spécificités de l'image ciblée. Certains observateurs seront ainsi souvent bien plus enclins à s'adapter, voire à outrepasser ces obscurations.

La robustesse des algorithmes générateurs comme DeepBlur, quant à elle, reste très bonne, que ce soit dans sa détection ou son obscuration. Cependant, cette technique, en plus d'avoir une fidélité visuelle très complexe à gérer, ne peut s'appliquer qu'à certaines utilisations précises. Comment faire, par exemple, si l'on veut cacher la présence d'un visage et non la simple identité de la personne ? Nous pensons que cela serait impossible avec la méthode actuelle de DeepBlur.

L'empoisonnement, quant à lui, possède une efficacité excellente lorsque les bonnes conditions sont remplies. Cependant, des algorithmes de détection de données empoisonnés peuvent réduire son efficacité. Dans le papier que nous avons cité, de nombreuses méthodes complexes ont été utilisées mais n'ont pas réussi à totalement éviter un aspect significatif sur le modèle. La robustesse de l'empoisonnement, contrairement aux autres méthodes, réside ainsi dans le volume. Tant que l'efficacité des algorithmes de détection n'est pas totale avec ~0% de faux positifs, la détérioration des CNN ciblés sera inévitable.

Importance du contexte et de l'objectif

Nous pensons que le plus important dans ces algorithmes est le contexte et l'objectif. De quel observateurs souhaite-t-on se protéger ? Pour quel format de données ? Quelles informations ou caractéristiques souhaitons-nous obscurcir ?

Toutes les approches pertinentes dans un contexte moderne prennent en compte chacun de ces concepts et composent au mieux leur approche afin de répondre au mieux à une certaine demande. Nous aurions également pu mentionner des algorithmes axés sur l'analyse de la convolution, qui peuvent devenir aussi spécifiques que le ciblage précis d'un CNN précis.

Conclusion

Notre approche au problème

A la lumière des avancées récentes des CNN et des approches d'obscuration, nous souhaitons explorer une approche hybride entre détérioration et empoisonnement. Pour cela, il faudra mettre en place un contexte d'entraînement et de test d'un ou plusieurs CNN. Notre objectif est de maintenir la meilleure fidélité visuelle possible, de façon rapide, tout en empêchant au maximum un CNN d'extraire des informations pertinentes sur l'image. De plus, nous souhaitons simuler des mécanismes de défense potentiels contre nos attaques.

Nos approches essaieraient ainsi à la fois d'empoisonner une partie des données d'entrainements du CNN et détériorer ses données de test. Nous imaginons plusieurs filtres possibles, qui s'attaquent à la catégorisation, aux données non pertinentes pour le système visuel humain ou même des obscurations naïves. Notre objectif est de tester un maximum de méthodes qui peuvent être pertinentes, les combiner, les attaquer, recommencer et en sortir une approche hybride efficace.

Annexes

Références

[1] - McPherson, R., Shokri, R., & Shmatikov, V. (2016). Defeating Image Obfuscation with Deep Learning. ArXiv, abs/1609.00408.

[2] - Li, T., & Choi, M. (2021). DeepBlur: A Simple and Effective Method for Natural Image Obfuscation. ArXiv, abs/2104.02655.

[3] - Shan, S., Ding, W., Passananti, J., Zheng, H., & Zhao, B.Y. (2023). Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. 2024 IEEE Symposium on Security and Privacy (SP), 807-825.

Dépot Github

- https://github.com/Patateon/Securite-visuelle