CEA List
Laboratoire de Vision et d'Ingénierie des Contenus
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Angélique LOESCH
Tél +33 (0)1 69 08 89 07
E-mail angelique.loesch@cea.fr

STAGE 2019 Réf : AS-19-S7

Vulnérabilité et défense des modèles de reconnaissance faciale par deep learning

Présentation du laboratoire d'accueil

Au cœur du Plateau de Saclay (Ile-de-France), l'institut CEA LIST focalise ses recherches sur les systèmes numériques intelligents. Porteurs d'enjeux économiques et sociétaux majeurs, ses programmes de R&D sont centrés sur les systèmes interactifs (intelligence ambiante), les systèmes embarqués (architectures, ingénierie logicielle et systèmes), les capteurs et le traitement du signal (contrôle industriel, santé, sécurité, métrologie).

Dédiés à la recherche technologique, les 700 ingénieurs-chercheurs et techniciens de l'institut ont pour objectif de favoriser l'innovation et son transfert autour de partenariats industriels pérennes. La culture projet et l'excellence scientifique des équipes de l'Institut sont au cœur de cette ambition.

Le Laboratoire de Vision et d'Ingénierie des Contenus mène ses recherches dans les domaines de la Vision par Ordinateur (Computer Vision) et l'analyse automatique de texte avec le défi d'extraire et d'organiser l'information à partir de documents faiblement ou non structurés (texte, image, vidéo, réseaux de capteurs).

Contexte du stage

Ce stage s'inscrit dans les activités d'Analyse de Scène du Laboratoire Vision et Ingénierie des Contenus du CEA List. Le laboratoire développe dans le cadre de projets R&D, des systèmes d'analyse automatique d'images et de vidéo intégrant des technologies avancées de détection, localisation, classification et suivi temporel d'objets d'intérêt, de reconnaissance d'activité et de comportements, ainsi que des outils de modélisation automatique de la scène.

Description du stage

La reconnaissance faciale par apprentissage automatique est aujourd'hui de plus en plus utilisée au quotidien comme solution d'authentification dans les smartphones ou encore d'identification de personnes dans le milieu de la vidéo surveillance. En quelques années, les systèmes proposés dans l'état de l'art sont même devenus capables de surpasser les performances humaines. Cependant, leurs modèles restent vulnérables : confrontés à des données d'entrée dites antagonistes (*adversarial examples*), ils échouent à reconnaître correctement une personne et présentent alors une faille de sécurité. Ces attaques sont d'autant plus dangereuses qu'elles sont souvent impossibles à distinguer pour l'homme. En général, elles sont en effet générées par des algorithmes appliquant aux données d'entrée de faibles perturbations (voir figure issue de [3]), mais toutefois suffisamment importantes pour obliger les modèles attaqués à produire de mauvaises prédictions [1][2][3][4]. Appliquées à la reconnaissance faciale (voir figure issue de [8]), ces attaques forcent un modèle à identifier à tort une personne comme correspondant à une autre et cela, en apportant d'infimes modifications à l'image originale. Ces données antagonistes posent dès lors de réels problèmes de sécurité, celles-ci pouvant être générées sciemment et utilisées pour attaquer les systèmes de reconnaissance faciale.

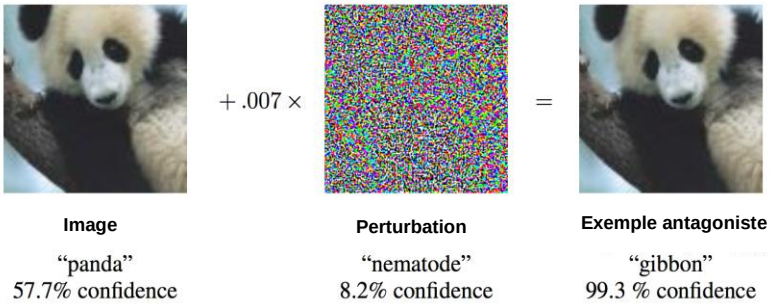

Illustration issue de [3] : Un exemple de génération d'exemple antagoniste appliqué au modèle classifieur GoogLeNet sur une image du jeu de données ImageNet.

CEA List
Laboratoire de Vision et d'Ingénierie des Contenus
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Angélique LOESCH
Tél +33 (0)1 69 08 89 07
E-mail angelique.loesch@cea.fr

La majorité de ces méthodes de génération d'attaques nécessite que l'opposant connaisse les paramètres du modèle de reconnaissance faciale pour résoudre le problème d'optimisation de la perturbation des images d'entrée. Cependant, certains travaux [6][7][8][9] considèrent comme plus réalistes les méthodes uniquement capables d'interagir avec le modèle attaqué en le considérant comme une boîte noire et en ne faisant qu'observer ses prédictions sur des images d'entrée spécifiques.

Lors de ce stage, on s'intéressera à ces méthodes de génération d'attaques par apprentissage profond en considérant le système comme une boîte noire. L'objectif sera dans un premier temps d'étudier les méthodes génératrices de l'état de l'art pour tromper les modèles de reconnaissance faciale par *deep learning*. Ces attaques antagonistes serviront à améliorer la robustesse des algorithmes de reconnaissance faciale. Puis dans un second temps, il sera intéressant de proposer d'autres types d'attaques toujours dans l'optique d'améliorer la défense de ces modèles.

Illustration issue de [8] : Un exemple de donnée antagoniste (G(Im1)) sur le visage Im1 tiré du challenge MCS 2018 « Adversarial Attacks on Black Box Face Recognition ». Sous l'effet de l'attaque, le modèle de reconnaissance automatique prédirait à tort que la personne Im1 et la personne Im2 serait identique.

Références

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[2] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

[3] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[4] Dong, Y., Liao, F., Pang, T., Su, H., Hu, X., Li, J., & Zhu, J. (2018). Boosting adversarial attacks with momentum. arXiv preprint.

[5] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In Security and Privacy, 2016 IEEE European Symposium on (pp. 372-387).

[6] Tang, D., Wang, X., & Zhang, K. (2018). Query-Free Attacks on Industry-Grade Face Recognition Systems under Resource Constraints. *arXiv preprint arXiv:1802.09900*.

[7] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (pp. 506-519). ACM.

[8] Milton, M. A. A. (2018). Evaluation of Momentum Diverse Input Iterative Fast Gradient Sign Method (M-DI2-FGSM) Based Attack Method on MCS 2018 Adversarial Attacks on Black Box Face Recognition System. *arXiv preprint arXiv:1806.08970*.

[9] Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.

Niveau demandé :	Ingénieur, Master 2
Ce stage ouvre la possibilité de poursuite en thèse et ingénieur R&D dans notre laboratoire.	
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.
Compétences requises : - Méthodes d'apprentissage automatique, deep learning - Tensorflow, python	