Adversarial attacks from a partially recovered network

Skyld, Marwane EL HANI, Victor GUYOMARD, Marie PAINDAVOINE

2 pages maximum / 2 pages maximum

Identification des cas d'usages / Identification of Use Cases

Contexte opérationnel. Les systèmes de défense embarquent des modèles d'apprentissages profonds de vision pour la détection et la classification de cibles, sur des porteurs contraints : drones, caméras de surveillance, calculateurs de bord. Pour des raisons de latence et d'autonomie énergétique, ces modèles sont souvent compilés en binaires optimisés pour le matériel cible (TensorRT pour NVIDIA Jetson, ou autres formats propriétaires équivalents). Ces modèles d'IA sont exposés: un porteur peut être abattu, capturé ou récupéré intact, et le ou les modèles associés analysé par un adversaire.
Problématique. L'évaluation de la robustesse adverse de ces systèmes repose aujourd'hui sur deux hypothèses extrêmes, qui encadrent mal le risque réel. L'hypothèse « boîte blanche » suppose l'adversaire en possession du modèle complet et dans un format permettant la différentiation : elle est pessimiste, car la compilation détruit une partie de la structure du réseau ainsi que les capacités de différentiation. L'hypothèse « boîte noire » suppose qu'il n'obtient rien du binaire et doit interroger le système des milliers de fois : elle est optimiste, car le binaire contient nécessairement l'ensemble des poids. La position réelle d'un adversaire ayant capturé un matériel embarqué se situe entre les deux, et n'est couverte par aucune méthode existante. Un donneur d'ordre ne dispose donc pas aujourd'hui d'un moyen de répondre à la question : *que peut réellement faire un adversaire ayant récupéré notre modèle compilé, optimisé ?*
Capacité à améliorer. Nous proposons de caractériser et d'outiller ce régime intermédiaire, afin de fournir une évaluation quantitative du risque de compromission via exemples contradictoires d'un modèle embarqué après capture du matériel.
Utilisateurs potentiels. Équipes d'évaluation et de qualification étatiques, laboratoires d'évaluation de la sécurité des produits, équipes red team des maîtres d'œuvre industriels, concepteurs de systèmes d'IA embarquée cherchant à dimensionner leurs contre-mesures.
Transposabilité civile. Le même schéma s'applique aux modèles embarqués dans les véhicules autonomes, les mobiles et les dispositifs médicaux connectés, où l'extraction de modèles depuis des paquets applicatifs est déjà documentée.
Expérimentation et passage à l'échelle. Le travail a été testé sur NVIDIA Jetson Orin avec TensorRT. Il a été évalué et validé sur ImageNet et des architectures de vision classiques.

Approche technique proposée / Proposed Technical Approach

Verrou traité. Un binaire compilé contient tous les poids du réseau, mais la fusion d'opérateurs et la réorganisation des tenseurs pour les noyaux de calcul rendent la reconstruction d'un modèle différentiable complet hors de portée. Les premières couches font exception, elles suivent des motifs standards, typiquement une convolution suivie d'une normalisation et d'une activation, que le compilateur fusionne en un opérateur identifiable. L'adversaire réaliste reconstruit donc un *préfixe* du réseau victime (gradients exacts jusqu'à une profondeur donnée, rien au-delà) tout en conservant la possibilité d'exécuter le binaire et d'en observer les sorties. Aucune attaque publiée n'exploite cette position : les méthodes boîte blanche exigent la tête de classification, les méthodes boîte noire ignorent les gradients disponibles et paient des milliers de requêtes par image.
Choix méthodologiques. Nous proposons une attaque adaptée à ce régime, en deux temps. L'objectif ne pouvant porter sur les logits, il est déplacé sur la sortie du préfixe : l'image source est optimisée pour que sa représentation intermédiaire se rapproche de celle d'une image réelle de la classe cible, appelée *guide*. Un guide unique étant peu fiable, l'attaque en utilise un petit ensemble et laisse le système déployé arbitrer, à raison d'une requête par candidat. La méthode ne nécessite aucun entraînement préalable et son coût en requêtes est exactement la taille de l'ensemble de guides.
Résultats de validation.
Nous arrivons en moyenne à reconstruire 3 à 6 couches du modèle compilé en format différentiable.

Sur ImageNet, avec un quart du réseau reconstruit et vingt requêtes par image, l'attaque atteint **70,5 % de succès ciblé** en moyenne sur quatre architectures (ResNet-50, Wide-ResNet-101-2, MobileNetV2, VGG-16), pour une perturbation bornée à 10/255. Le succès atteint 100 % dès que trois quarts du réseau sont reconstruits. Même un préfixe réduit à la première couche convolutive conserve un pouvoir de ciblage.

Technologies mobilisées. Rétro-ingénierie de moteurs d'inférence compilés (TensorRT), réimplémentation du préfixe dans un cadre à différentiation automatique (PyTorch), optimisation sous contrainte ℓ_∞..

Niveau de maturité. Prototype de recherche validé sur plusieurs exemples proche de la réalité.