La détection à coup sûr des deepfakes ne marchera sans doute jamais totalement

Mission impossible ?

6 min

Nicolas Beuve
Nicolas Beuve

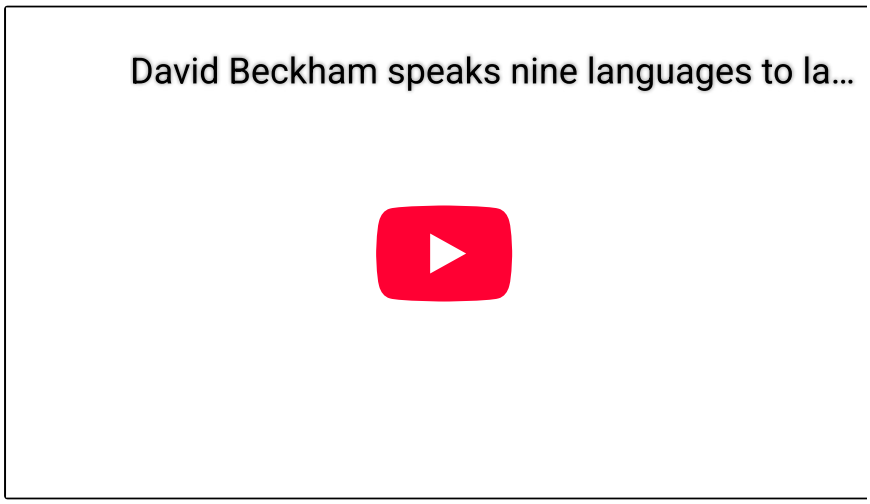
Résumer l'article Lecture Zen

Des systèmes d'intelligence artificielle créent des « deepfakes », ces vidéos ultra-réalistes, tandis que d'autres les détectent. Un jeu du chat et de la souris.

En mars 2022, environ un mois après l'invasion de l'Ukraine par la Russie, une vidéo de Volodymyr Zelensky, président de l'Ukraine, est diffusée sur une chaîne nationale ukrainienne. Dans cette vidéo, le président demande à son peuple de rendre les armes et de rentrer dans leurs familles. Cette vidéo a rapidement été identifiée comme un deepfake à cause de sa faible qualité et aura donc eu peu de répercussions sur les combats, mais cet exemple illustre parfaitement les dangers que peuvent poser les deepfakes.

Un deepfake est une vidéo dans laquelle le visage ou l'expression d'un individu a été volontairement modifié

permettant de modifier les expressions faciales ne datent pas d'hier, puisqu'il existe, par exemple, un article scientifique de 1999 utilisant des modèles 3D pour reconstruire et modifier des visages. Les deepfakes ont de nombreuses applications légitimes, par exemple dans le cinéma, la publicité et même plus récemment la compression vidéo.



Vidéo de David Beckham parlant neuf langues dans une vidéo pour *Malaria Must Die*.

Cependant, soutenus par l'essor du deep learning (une branche de l'intelligence artificielle dont le mot deepfake tire son nom), des modèles de génération de deepfakes grand public ont été développés. Ces modèles permettent à n'importe qui de créer gratuitement des deepfakes de bonne qualité, et ainsi d'usurper n'importe quelle identité dans des vidéos. Aujourd'hui, les deepfakes sont notamment utilisés

« Chat et moi sommes très proches ces derniers...

Google offre 12 mois de Gemini Pro et 2 To de stockage aux étudiants : comment e...

Google est-il derrière « cheetah », le modèle d'IA mystérieux apparu ce week-end ?

« À quel point suis-je foutu mec ? », il confesse ses actes de vandalisme à ChatGPT...

Quelles sont les meilleures alternatives gratuites à ChatGPT ?

Merci à nos Fondateurs

ipokamp Dam
Noisegratte Z
kennymat Yar

Retour

personnes de confiance est usurpée, poussant les victimes à effectuer des virements en pensant connaître leur interlocuteur.

S'il y a quelques années, une vidéo pouvait toujours être considérée comme authentique, ce n'est aujourd'hui plus le cas. On peut aussi se poser la question : peut-on détecter automatiquement ces fausses vidéos grâce à l'intelligence artificielle… alors qu'elles sont elles-mêmes générées par des IA ?

Comment modérer les deepfakes sur les réseaux sociaux ?

La modération des vidéos deepfake sur les réseaux sociaux est un sujet compliqué qui demande le développement de nouveaux outils.

Généralement, la modération des vidéos permet de filtrer les contenus violents ou haineux en utilisant des modèles d'intelligence artificielle spécifiquement entraînés à détecter ce genre de contenu. Cependant, dans le cas d'un deepfake, la vidéo peut être en apparence tout à fait inoffensive et ne pourra donc pas être détectée par ce genre de modèle. Par exemple, une vidéo de Volodymyr Zelensky demandant aux Ukrainiens de capituler n'a en apparence rien qui puisse justifier une suppression ; celle-ci ne devient un

Le deepfake de Volodymyr Zelensky a rapidement été supprimé des réseaux sociaux. // Source : Numerama

Dans les cas où les modèles d'intelligence artificielle ne sont pas capables de modérer les contenus, les réseaux sociaux s'appuient sur leurs utilisateurs ou des modérateurs humains pour filtrer le contenu. Mais encore une fois, ce genre de modération ne peut pas se transposer aux deepfakes, puisque les humains ne sont pas capables de détecter les deepfakes avec une grande précision.

Les outils de modération n'étant pas adaptés pour lutter contre les deepfakes, de nouveaux outils ont dû être développés, justifiant l'émergence d'un nouveau domaine de recherche : la détection automatique de deepfake.

Pour aller plus loin

Contre les deepfakes porno, le Sénat adopte des règles plus sévères

Première approche : mobiliser des détecteurs passifs

Les détecteurs sont généralement également basés sur le deep learning et peuvent être passifs ou actifs.

Retour

L'objectif d'un détecteur passif est de prédire si une image a été modifiée ou non sans connaître son origine. Un tel détecteur peut être utilisé par un utilisateur de réseau social, par exemple, pour tester la légitimité d'une vidéo en cas de doute.

Pour ce faire, le détecteur va utiliser des caractéristiques jugées comme discriminantes, c'est-à-dire qui permettent de distinguer facilement les deepfakes des images originales. Par exemple, en remarquant que les premiers deepfakes ne clignaient jamais des yeux, des chercheurs ont proposé de compter la fréquence des clignements. En dessous d'un certain seuil, les vidéos étaient alors étiquetées comme deepfakes.

Malheureusement, les modèles de génération se sont améliorés depuis, et ce genre de méthode n'est plus efficace, ce qui a poussé les chercheurs à développer de nouvelles techniques. Ainsi, la plupart des techniques récentes n'utilisent pas de connaissances d'experts, c'est-à-dire des caractéristiques choisies par l'homme, mais entraînent plutôt des modèles de deep learning, à l'aide de grandes bases de données contenant des vidéos étiquetées « real » ou « fake », à trouver leurs propres caractéristiques discriminantes.

Le problème majeur de ces techniques est qu'elles ne sont efficaces que pour détecter les deepfakes générés avec les méthodes utilisées pour constituer la base de données d'entraînement : elles ne se généralisent donc

Retour

de deep learning, bien que des propositions récentes commencent à y apporter des solutions.

Seconde solution : protéger l'image avant son détournement

À l'inverse d'un détecteur passif, un détecteur actif permet de protéger l'image originale avant que celle-ci ne soit modifiée. Ce genre de détecteurs est beaucoup moins populaire que les détecteurs passifs, mais pourrait notamment permettre aux journalistes de protéger leurs images, et ainsi d'éviter que celle-ci ne soit reprise à des fins de désinformation.

Une première méthode pour protéger une image consiste à y ajouter un filigrane, c'est-à-dire un message caché, qui peut être ensuite extrait de l'image. Ce message peut par exemple contenir des informations sur le contenu original de la vidéo, ce qui permet de comparer le contenu actuel de l'image avec celui caché. Dans le cas d'un deepfake, ces deux informations ne devraient pas concorder.

Des deepfakes de Tom Cruise // Source : Metaphysic.ai / TikTok

Une autre méthode de détection active consiste à appliquer une « attaque adverse » sur l'image afin d'empêcher la création de deepfakes. Une attaque adverse est une perturbation imperceptible d'une image, similaire à un filigrane, qui pousse un modèle de deep learning à l'erreur.

Par exemple, une attaque adverse ajoutée sur un panneau de signalisation « stop » pourrait perturber un modèle de deep learning qui n'y verrait pas un panneau-stop, mais par exemple une limitation à 130 km/h.

Dans le cas des deepfakes, une attaque adverse peut perturber le générateur de deepfakes et ainsi l'empêcher de produire un résultat de qualité.

Le problème des méthodes de détection actives est qu'elles modifient l'image et donc détériorent sa qualité, puisque les modifications appliquées aux pixels ne sont pas naturelles. De plus, les messages cachés étant de faible intensité, ils peuvent facilement être retirés, bien que leur suppression nécessite de réduire encore davantage la qualité de l'image, réduisant les chances que le deepfake soit confondu avec une image réelle.

Retour

Un jeu du chat et la souris sans fin ?

Le problème de la détection de deepfakes est encore aujourd'hui non résolu et risque malheureusement de ne jamais l'être complètement. En effet, les domaines de la détection et de la génération de deepfakes s'adaptent en permanence aux innovations de l'autre, avec un avantage évident pour la génération qui a toujours un coup d'avance sur la détection.

Ce constat ne signifie pas pour autant qu'investir dans la détection est une mauvaise idée. Les deepfakes disponibles sur Internet sont rarement issues des modèles de génération dernier cri et sont possiblement détectables. Dans le cas où un modèle de génération récent venait à être utilisé, la meilleure arme restera probablement de vérifier l'information véhiculée par la vidéo en utilisant différentes sources.

Nicolas Beuve, Doctorant en détection automatique de vidéo deepfake, *INSA Rennes*