Veille Cyber Jalon 3 : Création et Détection et du Deepfake

Valentin Thirion

Mai 2022

1 Abstract

Le deepfake est un processus qui permet de créer facilement de fausses images, vidéos ou sons à un tel niveau de réalisme qu'il est impossible de les discerner en tant qu'humain. Les méthodes de création du deepfake s'appuient sur des algorithmes de deep learning rendant la détection de ceux-ci quasi nulle via des méthodes de détection classiques. Pour palier ce problème des méthodes de détection utilisant des algorithmes de deep learning ont été créés. Ce document s'intéresse aux différentes méthodes de création d'un deepfake ainsi qu'aux méthodes de détection. Les problématiques qu'entraînent cette nouvelle technologie sont nombreuses (atteinte à l'image, diffusion de fausse informations, etc) et même si les moyens de détections s'améliorent, ceux de création aussi. C'est pourquoi il est nécessaire de prêter de plus en plus attention aux photos et vidéos que nous postons en ligne à titre individuel et de savoir repérer une fausse information.

2 Mots clés

Le domaine de l'intelligence artificielle peut sembler vaste et compliqué lorsqu'on ne connaît pas le vocabulaire associé. Voici les définitions de plusieurs mots clés essentiels à la bonne compréhension de ce document.

Intelligence Artificielle (IA): Systèmes ou machines qui imitent l'intelligence humaine. Une IA a généralement pour but de reproduire une faculté cognitive précise (comme le traitement du langage naturel, la reconnaissance d'images, etc). Ce type d'IA est dite faible car elle ne fait que reproduire un mode de fonctionnement humain, elle ne possède pas de conscience ou de sensibilité. Une IA capable de ressentir des sentiments et dotée d'une conscience est dite forte. A ce jour il n'existe que des IA faibles.

Machine Learning (ML) : Le machine learning (ou apprentissage automatique) est une branche de l'IA consistant à développer des systèmes qui apprennent ou améliorent leurs performances en fonction des données qu'ils traitent. Ils reposent sur des algorithmes utilisant des réseaux de neurones artificielles.

Deep Learning (DL) : Le deep learning (ou apprentissage profond) est une branche du ML qui a le même but mais qui ne repose pas sur les mêmes algorithmes. En effet ceux-ci utilisent des réseaux de neurones profonds.

Réseaux de neurones artificielles (ANN) : Un ANN est constitué de plusieurs neurones artificiels, chaque neurone est une fonction mathématique qui prend une ou plusieurs données x en entrée et sort une ou plusieurs données y en sortie. Un réseau peut être constitué de plusieurs couches de neurones et prendre des formes différentes. Un ANN est dit profond lorsqu'il possède plus de couches qu'un réseau classique (généralement plus de 4).

Réseaux de neurones convolutifs (CNN) : Type de réseau de neurones ayant pour but d'imiter des régions du cerveau spécifiques, notamment celles liées au champ visuel.

Réseaux de neurones récurrents (RNN) : Type de réseau de neurones adaptés pour l'analyse de séries temporelles (notamment pour la reconnaissance automatique de la voix ou l'analyse vidéo).

Modèle supervisé : Modèle qui nécessite l'aide de l'humain pour s'améliorer. Le modèle reçoit généralement une entrée accompagnée d'un label, ce label correspond au résultat que le modèle est censé retourner, si ce n'est pas le cas alors l'humain modifie le modèle jusqu'à obtenir un taux de réussite satisfesant

Modèle non-supervisé : Modèle qui apprend de lui-même sans aide humaine extérieure.

3 Introduction

Le deep learning est une branche de l'intelligence artificielle qui a permis de résoudre de nombreux problèmes complexes de nos jours. Allant des recommandations toujours plus poussées pour notre consommation jusqu'à la reconnaissance de maladies graves à des stades quasi invisible pour l'humain, le deep learning s'est imposée comme l'une des méthodes d'analyse des masses de données les plus utilisées. Cependant le deep learning est aussi une menace pour notre vie privée ainsi que pour la fiabilité des informations en ligne. En effet, une des utilisations du deep learning qui a récemment émergée est le deepfake. Les algorithmes de deepfake permettent de créer facilement de fausses

images, vidéos ou sons à un tel niveau de réalisme qu'il est impossible de les discerner en tant qu'humain ou même avec des méthodes de détection classiques.

Le deepfake existe depuis moins de 10 ans, c'est une technologie extrêmement récente en pleine évolution. Depuis sa création de nombreux trucages ont été réalisés, nous pouvons citer de fausses déclarations de guerre faite par l'ancien président Trump ou Obama en vidéo, de fausses images de documents classés secrets diffusés sur Twitter ou encore de faux appels téléphonique imitant la voix des dirigeants de grands groupes. Mais la première utilisation du deepfake reste l'incrustation de visages sur des vidéos à caractère pornographique. Ces exemples nous permettent de nous rendre compte des enjeux du deepfake et de la dangerosité de celui-ci s'il n'existe aucun moyen de le détecter.

Ce document a pour but de présenter les différentes méthodes de création d'un deepfake et les méthodes de détections développées pour les reconnaître. Il reviendra sur les différents enjeux que le deepfake entraîne et proposera une synthèse des problématiques. Enfin ce document conclura par une prédiction incertaine sur le futur de cette technologie et de quelques conseils sur l'hygiène informatique à tenir dans un monde où le deepfake est accessible à tous.

4 Présentation des sources de référence

Les sources de références peuvent être découpés en 3 parties : les méthodes de création des deepfake, les méthodes de détection des deepfake et les jeux de données servant à estimer si une nouvelle méthode de détection est fiable ou non. Les méthodes de création sont apparues en premier lieu, les méthodes de détection ont commencées à émerger après coup pour enfin laisser place aux jeux de données.

Les méthode de création peuvent être découpés en deux grandes familles : les réseaux antagonistes adverses ou Generative Adversarial Networks (GAN) et les doubles auto-encodeurs ou paire d'auto-encodeur.
Les méthodes de détection du deepfake en peuvent être classées en 4 grandes familles : les détections basées sur les fonctionnalités visuelles, locales, profondes et temporelles.

La présentation des sources suit un fil conducteur : la chronologie des innovations. C'est pourquoi nous présenterons d'abord les GAN puis les doubles auto-encodeurs avant de présenter les différentes méthodes de détections et les jeux de données. Voici un tableau récapitulatif des sources de références en fonction de leur année de parution, de leur type, du nombre de fois où elles ont été citées et du sujet qu'elles traitent.

Titre	Nombre de citations	Année	Type	Creation : GAN	Creation : Two Autoencoder	Detection : Visual feature-based	Detection : Local feature-based	Detection : Deep feature-based	Detection : Temporal feature-based	Dataset
[1]	41 398	2014	Article de colloque	X						
[2]	11 506	2014	Article de revue	X						
[3]	7 089	2014	Article de revue	X						
[4]	2137	2016	Article de revue		X					
[5]	381	2017	Article de revue		X					
[6]	281	2018	Article de colloque			X				
[7]	428	2019	Article de colloque			X				
[8]	255	2017	Article de colloque				X			
[9]	60	2018	Livre				X			
[10]	16	2019	Article de colloque				X			
[11]	171	2018	Article de colloque					X		
[12]	56	2018	Article de colloque					X		
[13]	506	2018	Article de colloque					X		
[14]	184	2019	Article de colloque					X		
[15]	489	2018	Article de colloque						X	
[16]	116	2019	Article de colloque						X	
[17]	244	2020	Article de colloque							X
[18]	106	2020	Article de colloque							X
[19]	159	2019 – 2022 (revisited)	Article de revue	X	X	X	X	X	X	X

Figure 1: Tableau comparatifs des sources de référence

4.1 Réseaux antagonistes adverses (GAN)

Un GAN est modèle de machine learning permettant de créer des données à partir d'exemples. Celui-ci peut être décomposé en deux choses distinctes : le générateur et le discriminateur. Le générateur est un modèle non supervisé, c'est-à-dire que l'humain donne des données x en entrée au générateur et celui-ci apprend de lui-même à reconnaître les caractéristiques importantes pour les reproduire. L'humain ne corrige pas le modèle, c'est le modèle qui se corrige lui-même.

Le discriminateur lui est un modèle supervisé, il a pour but d'identifier quelles sont les données originales et quelles sont les fausses données qu'il reçoit. Pour cela on lui fournit des données x en entrée mais ces données ont également un label y correspondant au résultat que le discriminateur est censé

obtenir, dans le cas d'un deepfake : "original" ou "faux". En fonction des résultats (appelés ici yhat) que donnera le discriminateur, un humain corrigera ou non le modèle. Ainsi plus le modèle s'entraîne plus il se perfectionne.

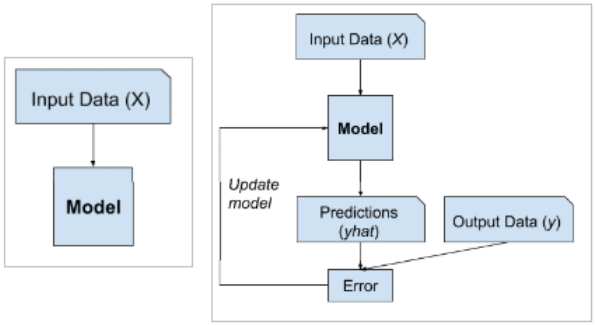

Figure 2: Modèle non-supervisé (gauche) et supervisé (droite)

En couplant le générateur et le discriminateur les deux modèles rentrent en compétition, ils deviennent adversaires (d'où le nom), l'un essaye de créer les meilleures fausses données, tandis que l'autre essaye de retrouver quelles sont les données originales. L'humain vérifie les résultats du discriminateur et change sa configuration si besoin. Ainsi les deux modèles deviennent de plus en plus performant jusqu'à obtenir un générateur capable de créer de fausses données imperceptibles. On déclare que le générateur est suffisamment performant lorsque le discriminateur se trompe une fois sur deux.

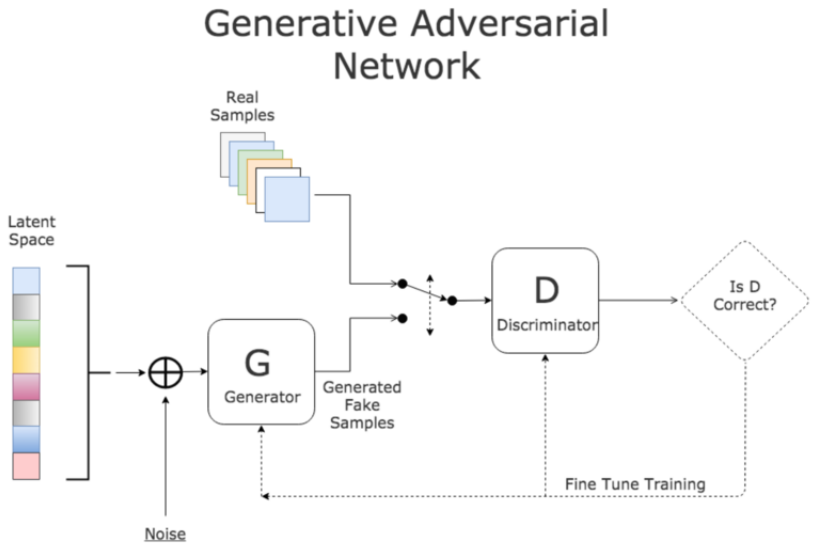

Figure 3: Schéma d'un réseau antagoniste adverse (GAN)

Les GAN ont été inventés par l'équipe de Ian Goodfellow [1], mais il faut bien comprendre qu'à la base ils n'étaient pas destinés à faire des deepfakes, le but était de créer de nouvelles données réalistes via un modèle très performant de machine learning que l'humain ne serai pas capable de produire car il n'aura pas forcément la même perception des caractéristiques importantes des données. Or il s'avère que si on prend des photos de visage comme données d'entrées, les GAN sont capables de créer de fausses personnes. Voir le fameux site : https://this-person-does-not-exist.com/fr

En revanche le premier modèle de GAN proposé comportait plusieurs problèmes, notamment deux : il nécessitait un humain pour entraîner le discriminateur, ce qui prenait beaucoup de temps. Et il ne permettait pas de faire de conditions, en effet dans certains cas le générateur pouvait créer des données mais ne percevait pas des caractéristiques très importantes des données en entrées, par exemple des oreilles pointues pour un chat. Il n'y avait aucun moyen de spécifier ces "conditions".
C'est pourquoi le modèle a très vite évolué, l'équipe de Radford [2] a réglé le problème de supervision et l'équipe de Mirza [3] le problème des conditions. Il faut tout de même noter que les GAN restent sujets à de nombreux problèmes, notamment la quantité de données pour entraîner le discriminateur et le temps avant d'obtenir un bon générateur.

4.2 Méthode par paire d'auto-encodeurs

La deuxième méthode pour créer des deepfake est la méthode par paire d'auto-encodeurs, celle-ci est considérée comme plus rapide et plus simple à mettre en œuvre, c'est d'ailleurs cette méthode qui aura été utilisée pour créer le premier deepfake de l'histoire d'Internet, en 2017 sur le forum Reddit.

Un auto-encodeur est composé de deux choses : un encodeur et un décodeur.
L'encodeur prend des données x en entrée et a pour but de les transformer en un modèle mathématique appelé "espace latent". Le décodeur prend cet espace latent en entrée et a pour but de retrouver au plus proche les données x qui ont été fournit à l'encodeur.
Les auto-encodeurs existent depuis très longtemps car c'est l'une des méthodes utilisées pour la compression de fichiers, l'encodeur prend un fichier volumineux en entrée et le traduit dans un modèle mathématique réduit. Le décodeur reprend ce modèle et reforme les données initiales du fichier.

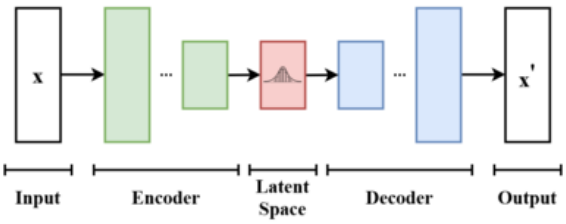

Figure 4: Schéma d'un auto-encodeur

La méthode par paire d'auto-encodeur utilise deux encodeurs identiques et deux décodeurs différents. Le premier auto-encodeur est entraîné pour reproduire un visage A tandis que le deuxième est entraîné pour un visage B. Une fois les deux modèles suffisamment entraînés il suffit de donner en entrée un visage A à un encodeur, celui-ci va alors le traduire en espace latent compréhensible pour les deux décodeurs, puis de donner cet espace latent au décodeur B, on obtient alors un deepfake car le décodeur B va essayer de reconstituer le visage de B à partir de l'espace latent du visage de A.

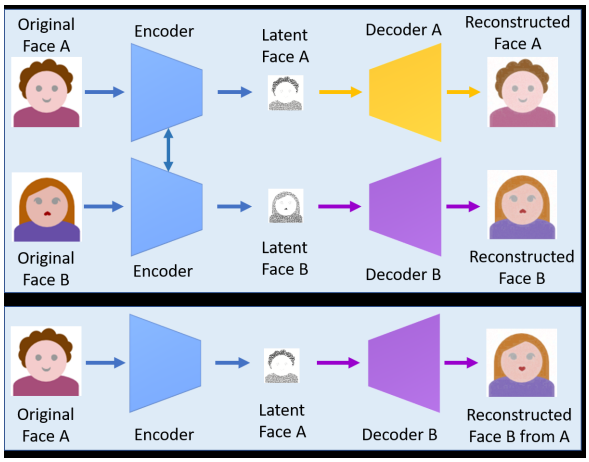

Figure 5: Schéma de la méthode par paire d'auto-encodeur

L'équipe qui a eu l'idée de mettre deux auto-encodeurs en parallèle est celle de Makhzani [4], tandis que l'équipe de Tewari [5] a utilisé ce modèle avec des visages pour la première fois.
Ce modèle est plus rapide et facile à mettre en place que les GAN car il nécessite moins de données en entrée mais il donne des résultats moins avancées en terme de deepfake, en effet ceux-ci sont plus simples à repérer via différentes méthodes de détection que nous allons aborder.

4.3 Détection par fonctionnalités visuelles

Les fonctionnalités visuelles sont des caractéristiques visibles à l'œil nu. Les fonctionnalités visuelles les plus utilisées pour détecter un deepfake sont : la différence de couleurs sur un même visage (avec de l'analyse de jeux d'ombres), le nombre de clignement des yeux et l'analyse de l'orientation du visage (pose de la tête, orientation des yeux, nez et bouche).

L'équipe de Haodong Li (non présente dans la bibliographie) utilise la première méthode et arrive à obtenir des résultats convaincants, cependant il est mentionné que certains GAN résolvent ces problèmes de couleurs et surtout d'ombres.

L'équipe de Yuezun Li [6] se base sur le fait qu'un humain cligne des yeux en moyenne une fois toutes les 6 secondes (entre 2 et 10 secondes à chaque intervalle) de manière inconsciente. Or les images qui entraînent les GAN sont très rarement des images de personnes qui clignent des yeux donc les GAN ne reproduisent que très peu ce mouvement. L'équipe à également obtenu de bons résultats mais considère que c'est un problème qui est facile à régler puisqu'il suffit de donner plus d'images de clignement des yeux aux GAN, de plus dans un cas de deepfake vidéo, si celui-ci peut être créé en prenant un visage (cible) et une vidéo en entrée, l'humain de la vidéo clignera des yeux de façon tout à fait normal ainsi que le deepfake résultant.

L'équipe de Xin Yang [7] traite l'orientation du visage par rapport aux yeux, nez et bouche via un modèle 3D reconstitué et arrive à déterminer si celui-ci est réel ou non. Leur méthode semble apporter de bons résultats mais n'a pas été évalué sur un grand nombre de tests.

4.4 Détection par fonctionnalités locales

Les fonctionnalités locales sont des caractéristiques qui ne sont visible à l'œil nu que dans certaines régions de la photo ou vidéo (par exemple un carré de 30 pixels par 20 pixels) mais généralement indiscernables sans l'aide d'un traitement informatique.

Les fonctionnalités locales les plus utilisés sont les bruits, ce sont de petites zones où les pixels ne sont pas "lisses", c'est-à-dire qu'ils ont des valeurs très différentes de la zone dans laquelle ils se trouvent, par exemple sur un t-shirt noir où tous les pixels devraient avoir des valeurs proches du noir et bien on peut trouver des pixels qui ont des valeurs proches du bleu-gris. Ces bruits sont des erreurs notamment formés dans la création de deepfake mais pas que.

Différentes équipes ont alors eu l'idée d'entraîner des modèles de réseaux de neurones convolutifs (CNN) pour détecter ces bruits, notamment dans les zones autour du visage qui sont susceptibles d'être plus marquées lorsqu'il s'agit de deepfake. Chaque équipe utilise un modèle différent, celle de Peng Zhou [8] utilise deux flux de CNN, tandis que celle de Marissa Koopman [9] et Zahid Akhtar [10] un seul.
Toutes les équipes se rejoignent sur un point, lorsque les deepfakes sont basés sur des photos ou vidéos de mauvaises qualités ces méthodes ne sont pas très efficaces car les bruits deviennent trop petits ou trop nombreux.

4.5 Détection par fonctionnalités profondes

Les fonctionnalités profondes sont des caractéristiques qui ne sont pas visibles à l'œil, elles reposent exclusivement sur le traitement informatique.

Les différentes équipes [11, 12, 13, 14] utilisent tous des modèles de deep learning pour leurs détections, certaines les entraînent via des réseaux sociaux [11], d'autres par des datasets [12, 13] et d'autres via différents médias comme des vidéos de journal TV [14].

Ces méthodes sont souvent bien plus longues et difficiles à mettre en place mais permettent de détecter plus de deepfake car le modèle reconnaît des caractéristiques que l'humain ne perçoit pas. Ces méthodes ont souvent de meilleurs résultats pour les photos et vidéos de mauvaises qualités, notamment celles sur les réseaux sociaux.

4.6 Détection par fonctionnalités temporelles

Enfin les fonctionnalités temporelles sont des caractéristiques qui sont visibles dans le temps, la détection de ces fonctionnalités n'est donc présente que sur des deepfake vidéos. Celles-ci extraient les caractéristiques de plusieurs images consécutives.

L'équipe de David Güera [15] récupère plusieurs dizaines d'images consécutives d'une vidéo et utilise des réseaux de neurones récurrents (RNN) pour déterminer si ce bout de vidéo est un deepfake ou non.

L'équipe de Irene Amerini [16] récupère également plusieurs images consécutives d'une vidéo et utilise des réseaux de neurones convolutifs pour analyser les flux optiques entre les différentes images. Si ceux-ci semblent irréels alors la vidéo est classée en tant que deepfake.

Ces méthodes de détection sont efficaces mais très longues à mettre en place car il faut d'abord entraîner un modèle puis analyser par très petits bouts la vidéo (une analyse peut prendre plusieurs heures pour quelques dizaines d'images consécutives), or dans une vidéo il peut y avoir plusieurs centaines de milliers d'images. Si la vidéo est truquée par deepfake à un endroit précis il faudra beaucoup de temps avant de le repérer.

4.7 Jeux de données

Les jeux de données sont des collections comportant des "samples" qui sont soit des deepfake soit des originales, ces samples peuvent être de différents types (photo, vidéo, son) et sont accompagnés d'un label pour préciser leur résultat.

Ces jeux de données peuvent comporter plusieurs niveaux de difficultés et plusieurs sections (comme "vidéos tirés de réseaux sociaux", "photos de mauvaises qualités", ...). Ils servent à évaluer la fiabilité des méthodes de détection.

Ainsi quand une nouvelle méthode de détection apparaît, celle-ci peut prouver son efficacité en montrant ses résultats sur différents jeux de données. L'équipe de Yuezun Li [17] et celle de Liming Jiang [18] ont crées deux jeux de données réputés et régulièrement utilisés.

Il faut également noter que les jeux de données sont souvent utilisés comme bases pour créer de bons GAN car les samples ont déjà un label, ainsi il est plus facile d'entraîner le discriminateur voir d'automatiser son apprentissage.

4.8 MindMap

Lorsque l'on parle de création ou de détection de deepfake il faut d'abord classer la nature du deepfake en question, il y en a trois possibles : vidéo / photo / son.

D'après les sources citées nous pouvons établir une MindMap regroupant les différentes méthodes de création et de détection par type de deepfake. Ce document n'a pas traité des deepfake sonores mais ceux-ci apparaissent dans la MindMap, la voici :

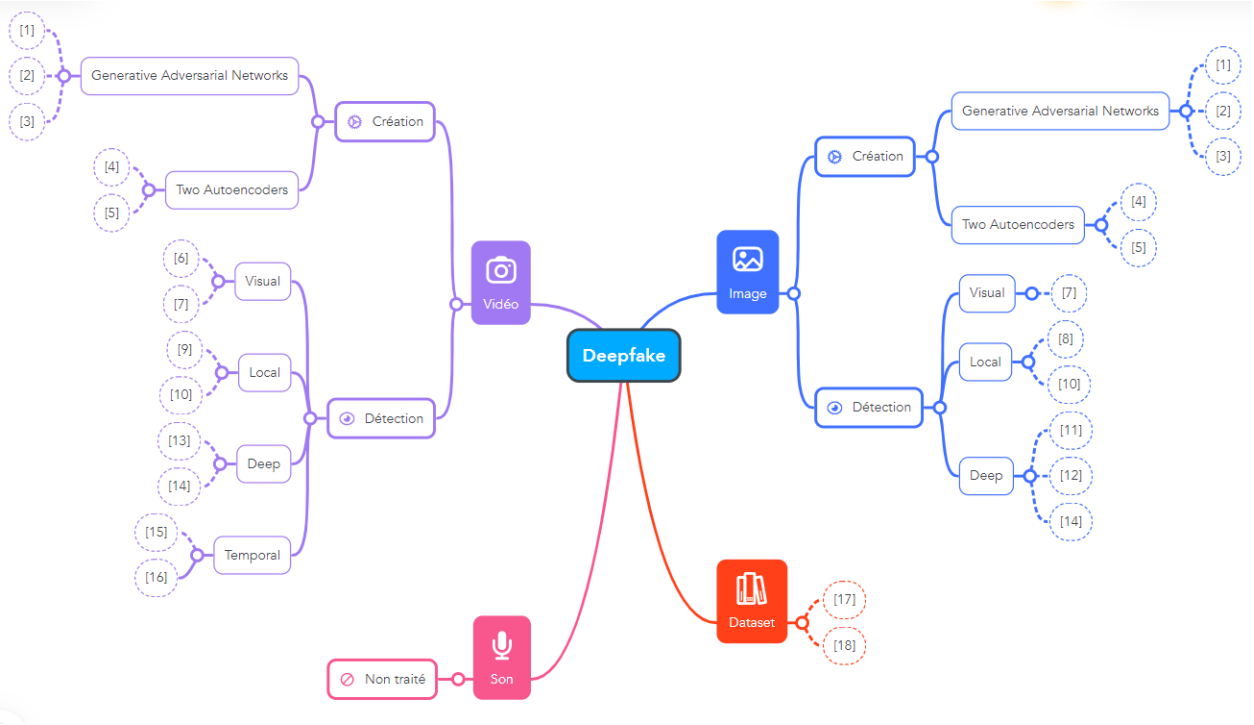

Figure 6: MindMap des différents thèmes abordés

5 Synthèse des problématiques

L'information est aujourd'hui considérée comme l'une des ressources les plus importantes si ce n'est la plus importante, or le deepfake est un moyen de créer ou de modifier à sa guise de fausses informations sans que personne ne puisse s'en apercevoir c'est pourquoi c'est l'un des sujets de recherches les plus importants actuellement.

Les enjeux sont majeurs et très larges, en effet le deepfake ayant énormément d'applications possibles, celui-ci engendre tout autant d'enjeux. D'abord la confiance générale vis-à-vis de l'information, en effet nous nous trouvons déjà à l'heure actuelle confronté à de nombreux problèmes via les "fakes news" qui sont de plus en plus nombreuses, ce phénomène de fausses informations circulant très vite couplé aux deepfake pourrait être dévastateur car il serait extrêmement compliqué d'identifier des sources fiables d'informations.

Ensuite les enjeux personnels, en effet la simple présence de photos ou de vidéos de votre identité sur internet pourrait permettre de vous mettre en scène et de vous faire dire n'importe quoi, c'est d'ailleurs la principale

utilisation du deepfake actuellement via la création de faux contenu haineux ou pornographiques.

Puis nous pouvons citer de nouveaux types d'attaques, notamment avec un mix d'ingénierie sociale et de deepfake, par exemple une attaque au président couplée avec une fausse vidéo ou une fausse voix générée par deepfake, rendrait l'authentification d'une personne au téléphone ou en visio-conférence très complexe. Ce type d'attaque a d'ailleurs déjà été observée dans de grands groupes et a parfois réussi grâce à des mécanismes psychologiques et au développement du télétravail.

Enfin tout ceci entraînerait des enjeux géopolitiques, en effet si la population ne peut plus distinguer quelles informations sont vraies de celles qui sont fausses alors certains pays pourraient être fragilisés via des élections influencées, des armées ne pouvant plus être sûrs de leurs ordres, des communications inter-étatiques non fiables, etc.

Il faut également noter que les générateurs utilisés dans les GAN sont à l'origine d'un autre modèle : les VQGAN. Ceux-ci permettent de créer une image totalement nouvelle à partir d'une phrase, entraînant alors de nombreuses problématiques dans le monde de l'art. En effet, des tableaux créés par un algorithme de deep learning (Dall-E 2) ont déjà été vendus. Si ce genre d'algorithme s'améliorent et se démocratisent cela voudrait dire que n'importe qui disposant d'une connexion Internet pourrait créer de l'art de façon illimité avec de simples phrases, rendant ainsi la limite flou entre le travail de "vrais artistes" et de particuliers.

6 Conclusion

Le processus de création et de détection du deepfake est encore très récent, celui-ci nécessite beaucoup de temps et surtout de données pour entraîner un modèle capable de fournir de bons résultats. Cependant nous avons vu que les méthodes de création deviennent de plus en plus rapides et simples à mettre en place, nous pouvons citer de nombreuses applications comme FaceApp ou SnapChat mettant à dispositions des filtres reposant sur le deepfake. Cette technologie devient donc de plus en plus accessible là où les méthodes de détection peinent à suivre la cadence.

L'enjeu de développer des méthodes de détection fiables est donc l'un des plus importants, c'est d'ailleurs la raison pour laquelle de grands groupes comme Apple, Meta et Google financent de nombreuses équipes de recherches à l'heure actuelle. Pourtant, comme nous l'avons vu dans les présentations la plupart

des méthodes de détection actuelles possèdent des faiblesses et ne permettent pas aujourd'hui de se défendre de façon simple et rapide contre les deepfake.

Nous pouvons nous attendre à voir surgir de plus en plus d'applications puissantes proposant des services de création de deepfake. Le partage excessif de données personnelles, notamment de photos et vidéos permettra d'alimenter les jeux de données avec lesquels ces algorithmes s'entraînent et ainsi de les rendre plus performants. La tendance de partager le plus rapidement possible l'information plutôt que de prendre le temps de la vérifier est aussi un vecteur aggravant. C'est pourquoi le travail de recherche et surtout de sensibilisation doit être amélioré sur ce sujet dans les années à venir.

7 Bibliographie

[1] I. Goodfellow *et al.*, « Generative Adversarial Nets », in *Advances in Neural Information Processing Systems*, 2014, vol. 27. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[2] A. Radford, L. Metz, et S. Chintala, « Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks », 2015, http://arxiv.org/abs/1511.06434

[3] M. Mirza et S. Osindero, « Conditional Generative Adversarial Nets », *arXiv:1411.1784 [cs, stat]*, nov. 2014, http://arxiv.org/abs/1411.1784

[4] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, et B. Frey, « Adversarial Autoencoders », *arXiv:1511.05644 [cs]*, mai 2016, http://arxiv.org/abs/1511.05644

[5] A. Tewari *et al.*, « MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction », p. 10.

[6] Y. Li, M.-C. Chang, et S. Lyu, « In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking », in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, déc. 2018, p. 1-7. doi: 10.1109/WIFS.2018.8630787.

[7] X. Yang, Y. Li, et S. Lyu, « Exposing Deep Fakes Using Inconsistent Head Poses », in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mai 2019, p. 8261-8265. doi: 10.1109/ICASSP.2019.8683164.

[8] P. Zhou, X. Han, V. I. Morariu, et L. S. Davis, « Two-Stream Neural Networks for Tampered Face Detection », in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, juill. 2017, p. 1831-1839. doi: 10.1109/CVPRW.2017.229.

[9] M. Koopman, A. Macarulla Rodriguez, et Z. Geradts, *Detection of Deepfake Video Manipulation*. 2018.

[10] Z. Akhtar et D. Dasgupta, « A Comparative Evaluation of Local Feature Descriptors for DeepFakes Detection », in *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, nov. 2019, p. 1-5. doi: 10.1109/HST47167.2019.9033005.

[11] F. Marra, D. Gragnaniello, D. Cozzolino, et L. Verdoliva, « Detection of GAN-Generated Fake Images over Social Networks », in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, avr. 2018, p. 384-389. doi: 10.1109/MIPR.2018.00084.

[12] C.-C. Hsu, C.-Y. Lee, et Y.-X. Zhuang, « Learning to Detect Fake Face Images in the Wild », in *2018 International Symposium on Computer, Consumer and Control (IS3C)*, déc. 2018, p. 388-391. doi: 10.1109/IS3C.2018.00104.

[13] D. Afchar, V. Nozick, J. Yamagishi, et I. Echizen, « MesoNet: a Compact Facial Video Forgery Detection Network », in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, déc. 2018, p. 1-7. doi: 10.1109/WIFS.2018.8630761.

[14] H. H. Nguyen, F. Fang, J. Yamagishi, et I. Echizen, « Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos », *arXiv:1906.06876 [cs]*, juin 2019, http://arxiv.org/abs/1906.06876

[15] D. Güera et E. J. Delp, « Deepfake Video Detection Using Recurrent Neural Networks », in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, nov. 2018, p. 1-6. doi: 10.1109/AVSS.2018.8639163.

[16] I. Amerini, L. Galteri, R. Caldelli, et A. Del Bimbo, « Deepfake Video Detection through Optical Flow Based CNN », in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), oct. 2019, p. 1205-1207. doi: 10.1109/ICCVW.2019.00152.

[17] Y. Li, X. Yang, P. Sun, H. Qi, et S. Lyu, « Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics », in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, juin 2020, p. 3204-3213. doi: 10.1109/CVPR42600.2020.00327.

[18] L. Jiang, R. Li, W. Wu, C. Qian, et C. C. Loy, « DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection », in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, juin 2020, p. 2886-2895. doi: 10.1109/CVPR42600.2020.00296.

[19] T. T. Nguyen *et al.*, « Deep Learning for Deepfakes Creation and Detection: A Survey », *arXiv:1909.11573 [cs, eess]*, févr. 2022, http://arxiv.org/abs/1909.11573.