Synthèse de la présentation IA générative et deepfake

Loric Gros, Mathys Sambet, Matéo Munoz

Octobre 2024

1 L'IA générative

L'IA générative, un type d'intelligence artificielle capable de créer du contenu multimédia comme des textes, images ou vidéos, transforme de nombreux secteurs en répondant aux besoins d'innovation et d'efficacité. Son fonctionnement repose sur différents modèles, tels que les Autoencoders Variationnels (VAE), les réseaux antagonistes génératifs (GAN) et les Transformers. Les VAE compriment les données pour générer du contenu, les GAN fonctionnent par duel entre un générateur et un discriminateur pour produire des contenus de haute qualité, et les Transformers, utilisés par des systèmes comme ChatGPT, permettent de traiter rapidement des séquences textuelles pour des applications comme la traduction ou la création de texte.

Les utilisations de l'IA générative sont diverses et étendues. Elle stimule la créativité en permettant la création d'art visuel et de musique, et elle améliore l'accessibilité aux connaissances en rendant l'information plus abordable et en facilitant des traductions et résumés. Dans le secteur professionnel, elle accroît l'efficacité en automatisant les tâches répétitives, tandis que dans le domaine de l'éducation, elle soutient l'apprentissage interactif et le développement de compétences personnalisées. Ces usages favorisent également la collaboration en permettant des échanges créatifs sur des plateformes collaboratives.

Cependant, l'IA générative comporte des défis importants. Elle peut être source de désinformation et de fausses informations, menaçant la vie privée et posant des risques pour l'emploi. Les questions éthiques autour du droit d'auteur et de la valeur de la créativité humaine deviennent cruciales, d'autant plus que la production de contenu avec ces systèmes a une empreinte carbone significative. Face à ces enjeux, la réglementation et la sensibilisation du public sont essentielles pour assurer une utilisation éthique et responsable de l'IA générative.

2 Risque et enjeux

Historiquement, la manipulation de média nécessitait des compétences techniques avancées, ce qui limitait les abus. Avec l'arrivée des deepfakes (2014), les GAN pouvait générer des images de faible qualité. Les progrès technologiques (2017) ont rendu ces contenus réalistes et accessibles, augmentant les risques de fraude.

L'utilisation de l'IA générative entraîne une déshumanisation des relations, les interactions sont plus le résultat d'un calcul et manque d'authenticité. Cela risque d'impacter l'éducation et de remplacer les échanges réels qui conduirait à un isolement social et à des effets émotionnels négatifs.

Les deepfakes très réalistes facilitent la propagation de fausses informations et influencent l'opinion publique, surtout via les réseaux sociaux où elles se diffusent rapidement. Même si un démenti est publié, il reste souvent moins vu. Cela peut aussi réduire la confiance dans les informations vraies, les gens risquant de douter de tout contenu visuel.

Les deepfakes peuvent créer des vidéos diffamatoires menaçant la vie personnelle et professionnelle, posant des problèmes de droit à l'image. Ils présentent aussi des risques de sécurité, notamment par l'imitation de voix et de visages, comme dans un cas de faux enlèvement réalisé via une voix simulée.

La créativité humaine est menacée, forçant les artistes à s'adapter aux technologies d'IA. Par exemple, le

photographe Boris Eldagsen a gagné un concours avec une image générée par IA, illustrant la difficulté de distinguer ces créations des œuvres humaines.

3 Quelques solutions pour contrer les deepfakes

En France, des articles de loi, tels que les articles 226-1 et 226-8 du Code pénal, punissent l'utilisation non consentie de l'image d'une personne, avec des sanctions allant jusqu'à 45 000 € d'amende et deux ans de prison.

Pour protéger les images, des tatouages numériques peuvent être utilisés. Ils modifient certains pixels de façon invisible pour empêcher leur utilisation dans les algorithmes d'apprentissage, protégeant ainsi les œuvres artistiques et données sensibles.

Il existe divers algorithmes pour détecter les deepfakes, comme celui basé sur le réseau de neurones VGG-16, composé de 13 couches de convolution, de max pooling et 3 couches fully-conected, qui réduit la taille des données tous les 2 à 3 convolutions. VGG-16, souvent utilisé pour la reconnaissance faciale, peut être combiné avec le modèle ResNet, dont les couches sautent certaines connexions lorsque les valeurs sont proches de zéro. Appliqué à la détection du clignement des yeux, ce duo d'algorithmes permet une précision de détection de 93 % à 98 %.

Une autre méthode de détection de deepfakes utilise l'apprentissage par "triple loss". D'abord, le modèle MTCNN identifie les zones d'intérêt et affine les contours des visages (par 3 phases de plus en plus précise). Ensuite, le modèle Xception, une amélioration de VGG-16, applique des convolutions sur chaque canal d'entrée pour extraire des détails précis des visages. Avec l'apprentissage "triple loss", un visage "anchor" et un visage "positif" (réels) sont rapprochés, tandis qu'un visage "négatif" (deepfake) est éloigné, ce qui permet d'atteindre une précision de détection jusqu'à 99,7 %.

4 Lien avec la cognition

La reproduction des biais cognitifs est présente dans les IA génératives qui vont avoir tendance à représenter ces biais sociaux dans leur création. Un autre point est l'auto-amélioration sans supervision qui est réalisée dans le fonctionnement des GAN. Un début pour l'embodiment est possible avec la création des voix et visages qui pourraient simuler l'interaction humaine.

5 Résumé des débats

Pensez-vous qu'un jour, il deviendra impossible pour un être humain de discerner les deepfakes sans assistance technologique ? Que cela implique-t-il ?

Cela entraîne une dépendance aux outils de détection, un biais de confiance envers les contenus visuels et une course entre faussaires et technologies de vérification. Pour garantir l'authenticité, des solutions comme la signature numérique et une responsabilité accrue des plateformes seront cruciales, bien que cela soulève des questions éthiques et logistiques.

Qui doit apprendre à la population à se servir des IA génératives ?

L'apprentissage de l'utilisation des IA génératives devrait être pris en charge par l'État et les institutions éducatives, car l'IA est un outil puissant nécessitant une compréhension de ses usages responsables et des risques associés. Une sensibilisation pourrait inclure des formations pratiques, encadrées par des lois adaptées et, si nécessaire, des taxes pour limiter les usages abusifs, tout en promouvant des pratiques éthiques et sécurisées.

Comment croyez vous que l'IA générative dans sa globalité sera utilisée dans le futur ?

Dans le futur, l'IA générative sera probablement utilisée de façon étendue dans les domaines créatifs, mais elle devra être encadrée pour protéger les artistes et les créateurs. La question du remplacement

des artistes est délicate car leur travail est profondément lié à l'originalité et à l'expression humaine. Un cadre devra être mis en place pour garantir le droit à la voix des professionnels comme les doubleurs et éviter que ceux qui souhaitent des œuvres sans IA ne soient contraints d'y être exposés. La régulation pourrait inclure une identification claire des créations IA et des limites pour les entreprises.