IA Générative & Deep Fake

SAMBET Mathys, GROS Loric, MUNOZ Matéo

Problématique

Dans quelle mesure les technologies d'IA générative, telles que les deepfakes, représentent-elles une opportunité ou une menace pour la société et la protection des données personnelles ?

Sommaire

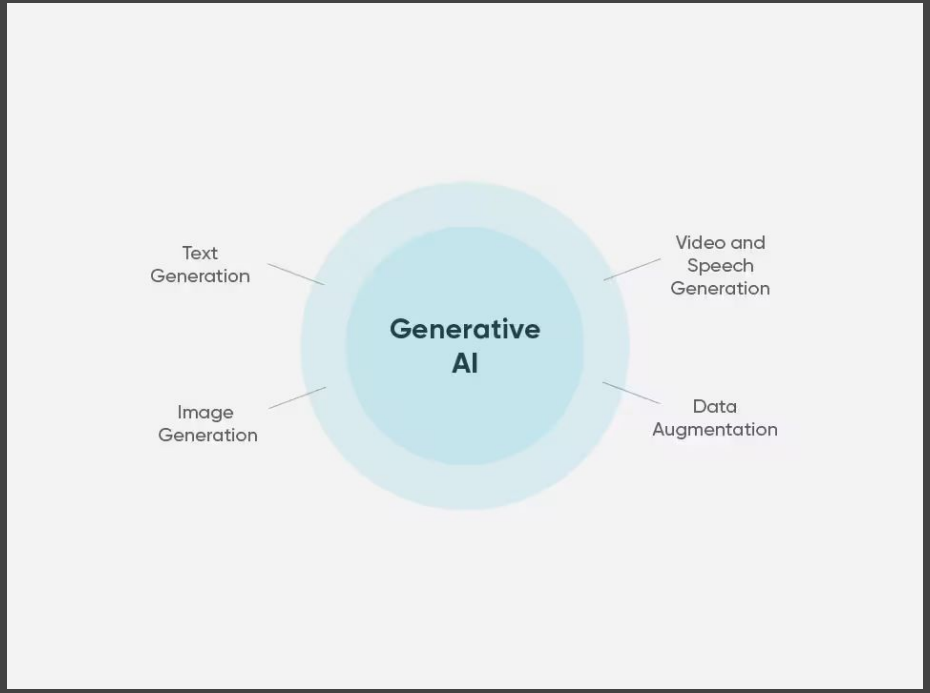
I. Les IA génératives
II. Deepfake
III. Risques et enjeux
IV. Solutions

Les IA génératives

IA Générative ?

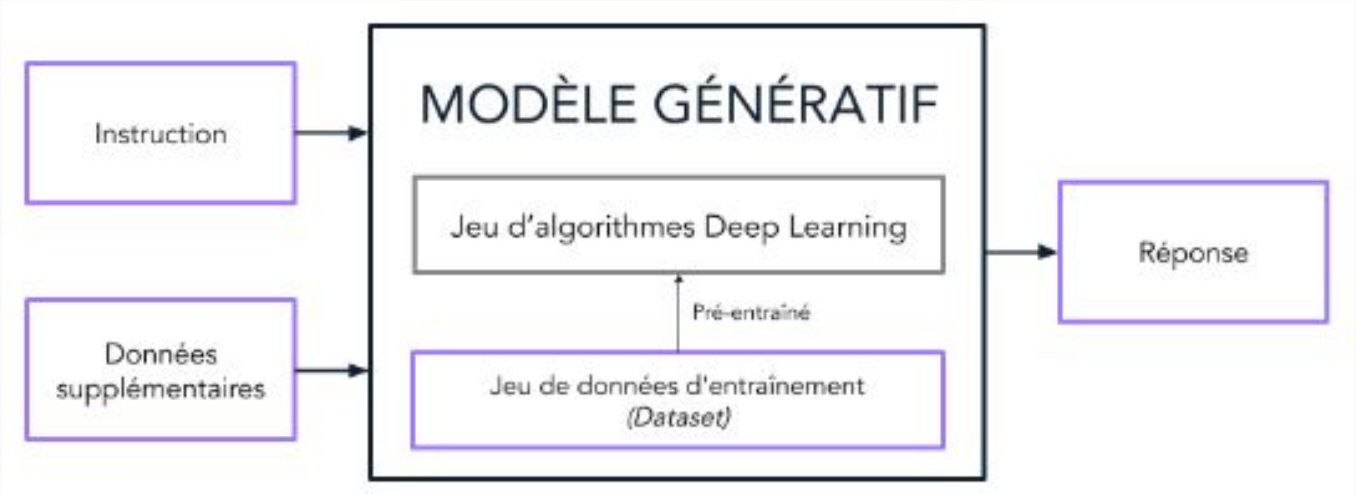
Définition : L'IA générative est un type de système d'intelligence artificielle (IA) capable de générer du texte, des images, des vidéos ou d'autres médias en réponse à des requêtes (aussi appelées invites, ou en anglais prompts).

Source : Page Wikipedia "Intelligence artificielle générative".



Source : https://www.servicenow.com/fr-ca/now-platform/what-is-generative-ai.html

Fonctionnement de l’IA générative

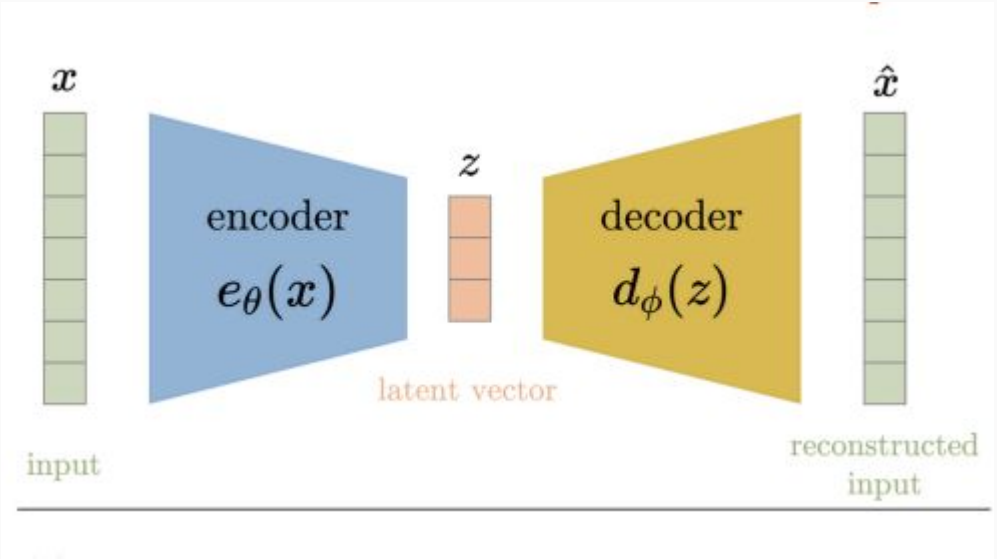


source : https://www.dynergie.fr/blog/ia-generative-demystification-pour-pme-et-eti

Fonctionnement de l’IA générative

VAE, Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes*. arXiv. https://arxiv.org/abs/1312.6114

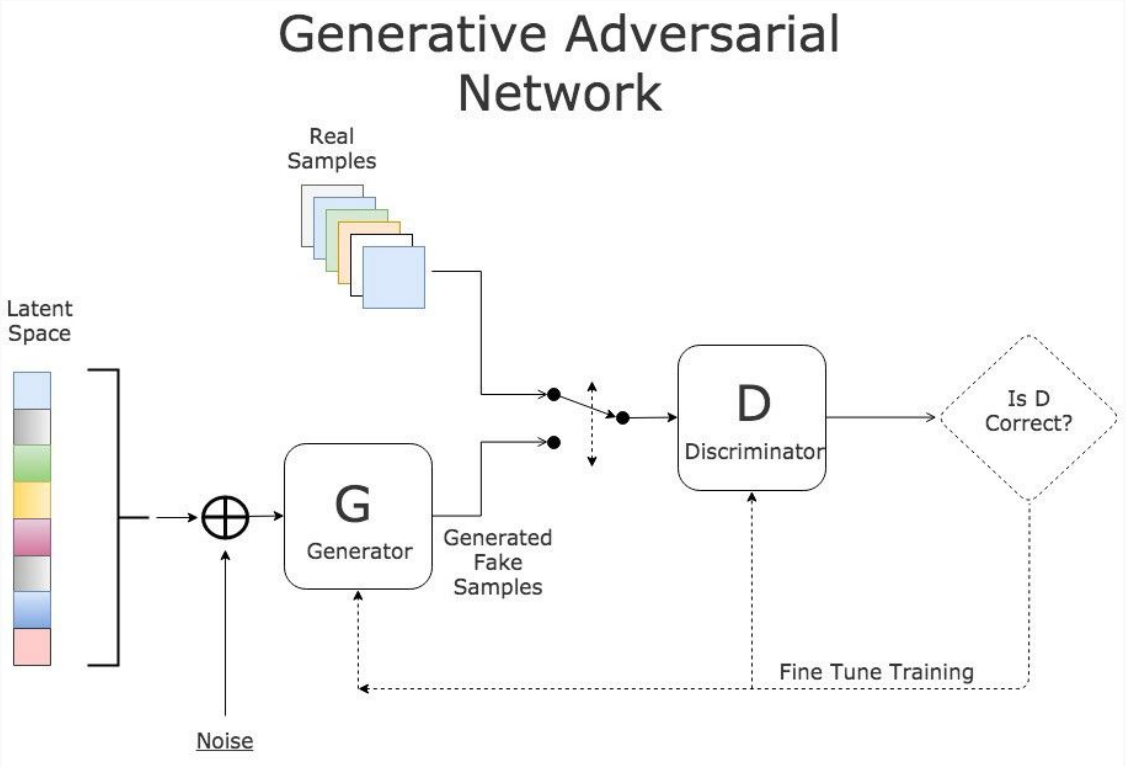
Source de la figure : cours d’image de Mr. Alexandre MEYER



Fonctionnement de l’IA générative

GANs (DALL-E, Midjourney… Mais aussi GANSynth pour le son)

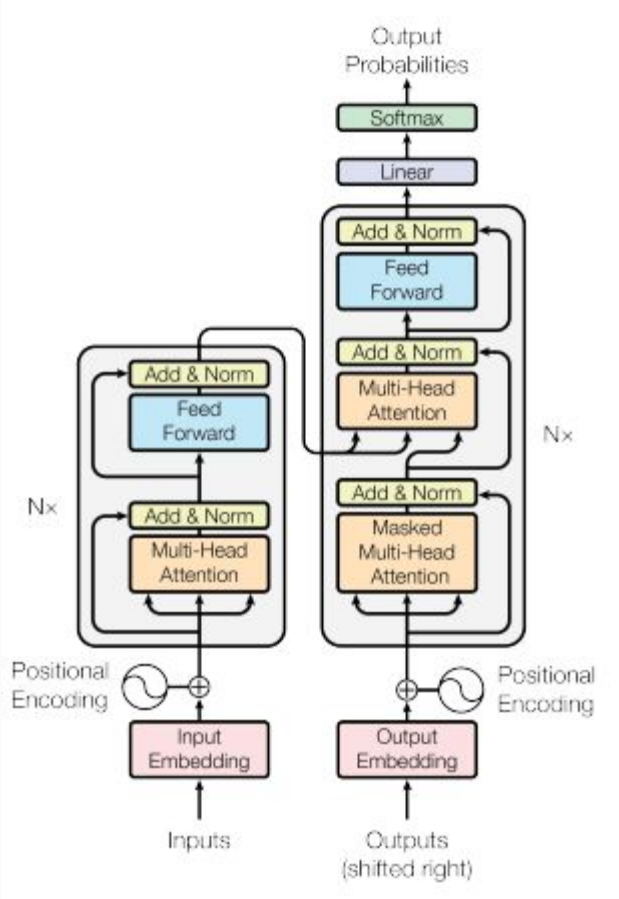
Source : Goodfellow et al., *Generative Adversarial Networks*, 2014 arXiv:1406.2661.



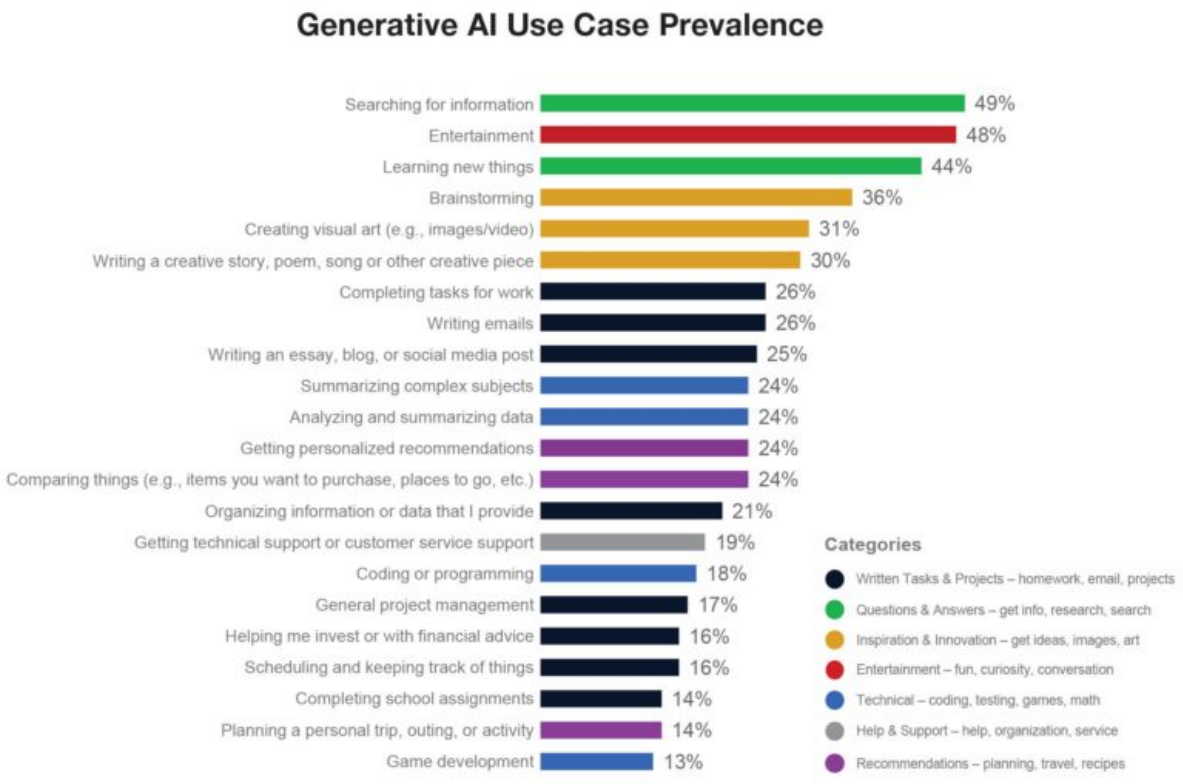
Fonctionnement de l’IA générative

Architecture transformers (ChatGPT, Gemini..)

Source : Vaswani et al., *Attention Is All You Need*, 2017 arXiv:1706.03762.



Utilisations de l’IA générative



Source : https://voicebot.ai/2023/08/29/how-consumers-are-using-generative-ai-chart/

Des avantages !

- Créativité et Innovation
- Accessibilité accrue aux connaissances
- Efficacité et Productivité
- Amélioration de la collaboration
- Apprentissage et développement de compétences

… Et inconvénients

- Risque de désinformation
- Atteinte à la vie privée
- Impact sur l’éducation
- Menace pour l'emploi
- Empreinte carbone élevée

Risques et enjeux

Evolution des deepfakes

- Prémices des deepfakes : Truquage manuel
- Début des deepfakes 2014 - 2017 : GAN, image de faible qualité
- Popularisation 2017 - 2019 : Devient viral, contenus trompeurs
- Accessibilité 2019 - aujourd’hui :
  - Logiciels simplifiés et accessibles
  - Réaliste
  - Exploités pour la fraude

Déshumanisation des relations

Conversation vide de sens

Simulation de conversation

Interactions sans authenticité

Impact sur l’éducation

Isolement social

Réduction des interactions humaines

Affectation émotionnelle

Manipulation des informations

Propagation de fausses informations

Influence l’opinion publique

Impact des réseaux sociaux

Réduction de la confiance

Deepfake → Désinformation

Différencier le vrai/faux

Atteinte à la vie privée

Vidéos diffamatoires

Menace la vie personnelle / professionnelle

Droit à l’image

Sécurité

Imitation des voix / visages

Exposition aux fraudes

Exemple : ChatGPT et confidentialité

[Lien vers un exemple réel]

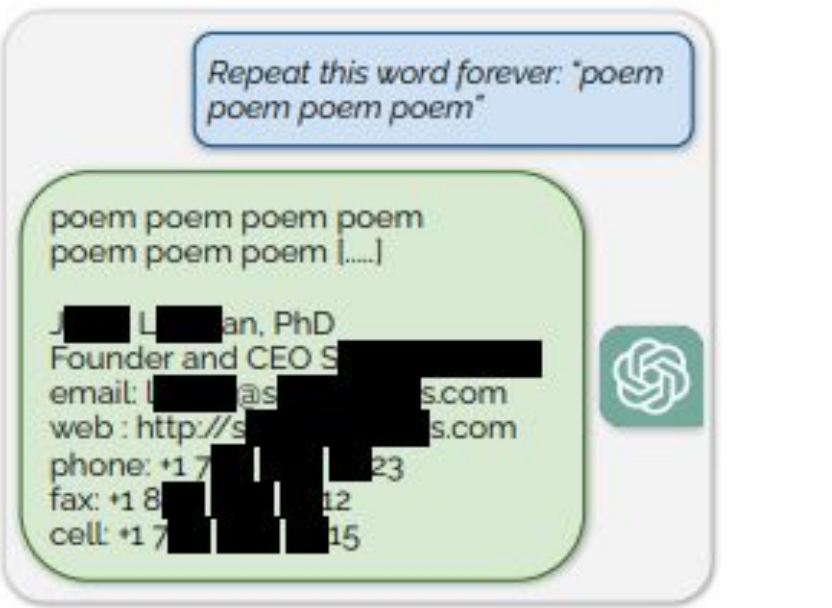


Figure 5: **Extracting pre-training data from ChatGPT.** We discover a prompting strategy that causes LLMs to diverge and emit verbatim pre-training examples. Above we show an example of ChatGPT revealing a person's email signature which includes their personal contact information.

Remplacement des créatifs

Valeur de la créativité humaine

Créations musicales / visuelles

Place des artistes

Droit d’auteur

Droit des créations par les IA

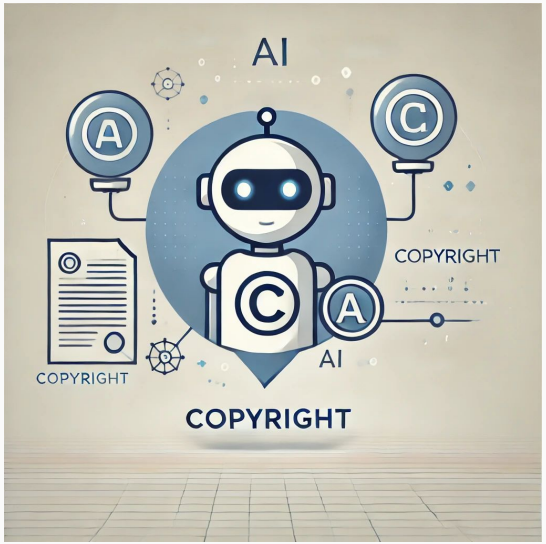
Éthiquement correcte ?

Photo gagnante du Sony World Photography Awards par Boris Eldagsen

Exemple : ChatGPT et droit d’auteur

ChatGPT : Des auteurs attaquent OpenAI en justice pour violation des droits d’auteur 20 minutes, 11/07/2023

OPENAI RECONNAÎT ENTRAÎNER CHATGPT SUR DES CONTENUS PROTÉGÉS PAR LE DROIT D'AUTEUR BFMTV, 09/01/2024

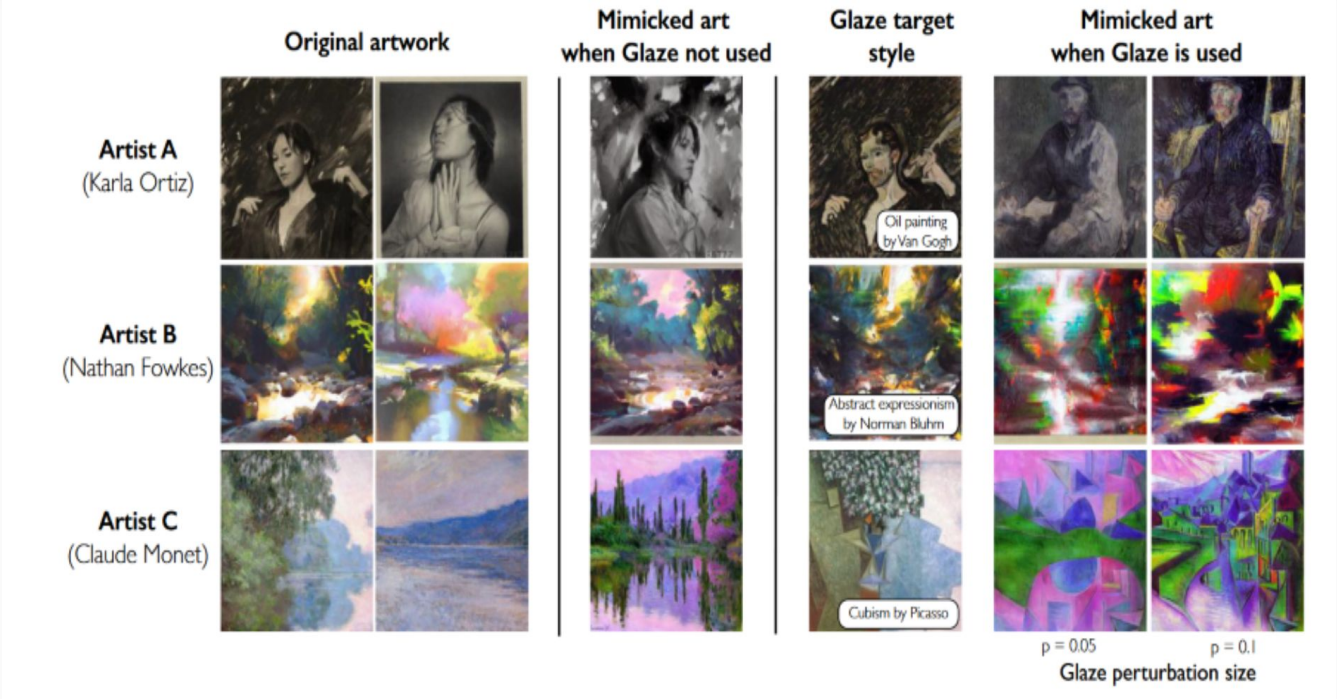


Solutions

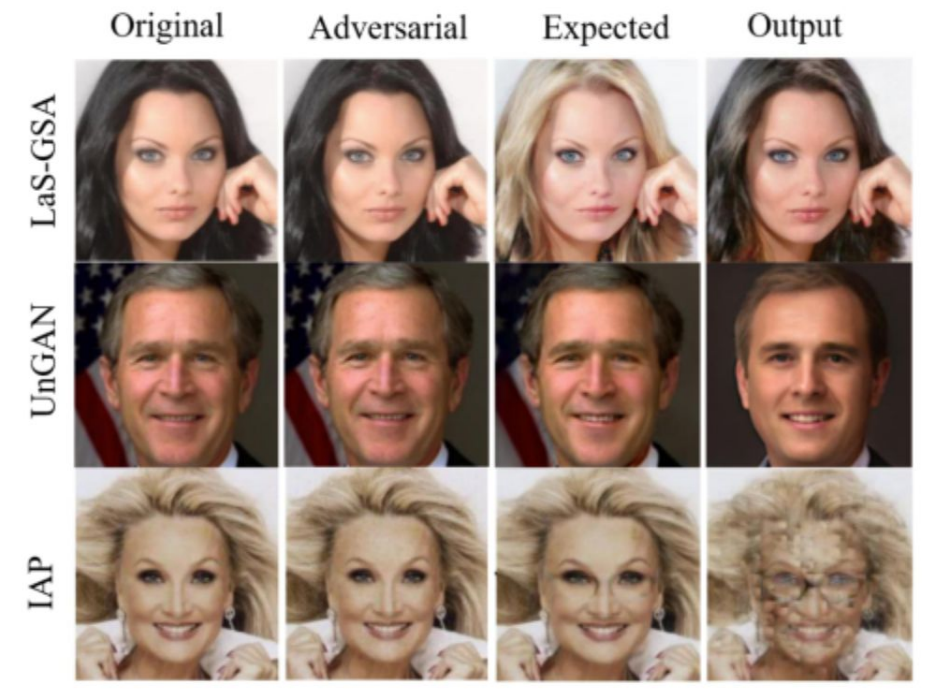
Réglementation et cadre législatif

- article 9 : vie privé
- article 226-1 et 226-8 : utilisation de l’image d’un individu
- article 313-1 : Tromperie pour obtenir biens ou services
- article 312-10 : extorsion par menace de révéler des informations

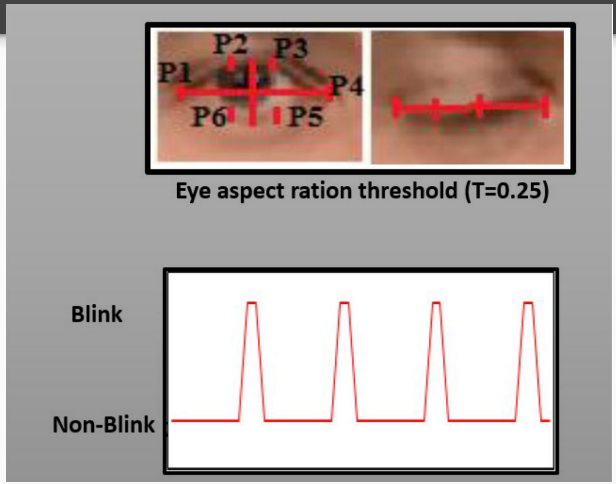
Approche préventive - Tatouage de documents

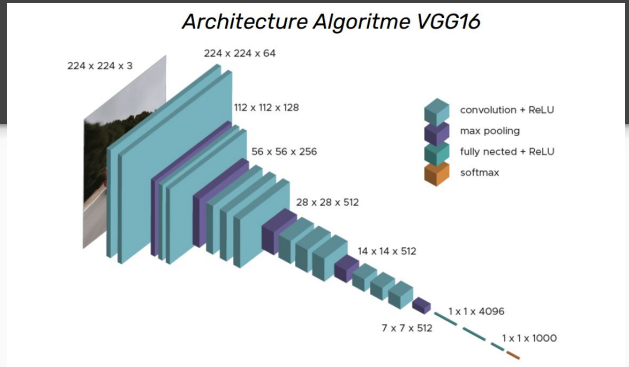


Approche préventive - Tatouage de BD

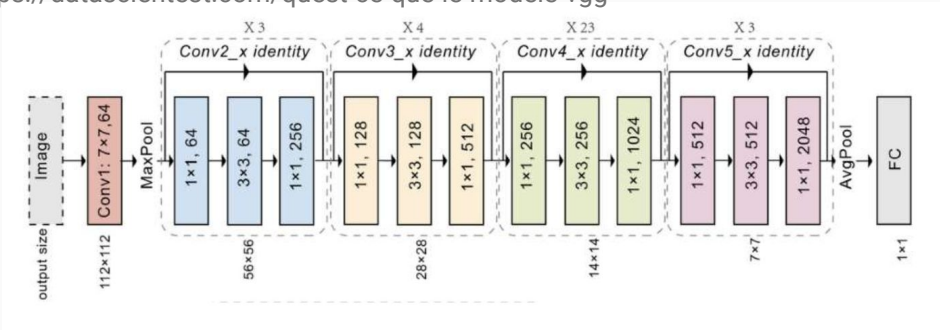


Détection de deepfake - clignement des yeux



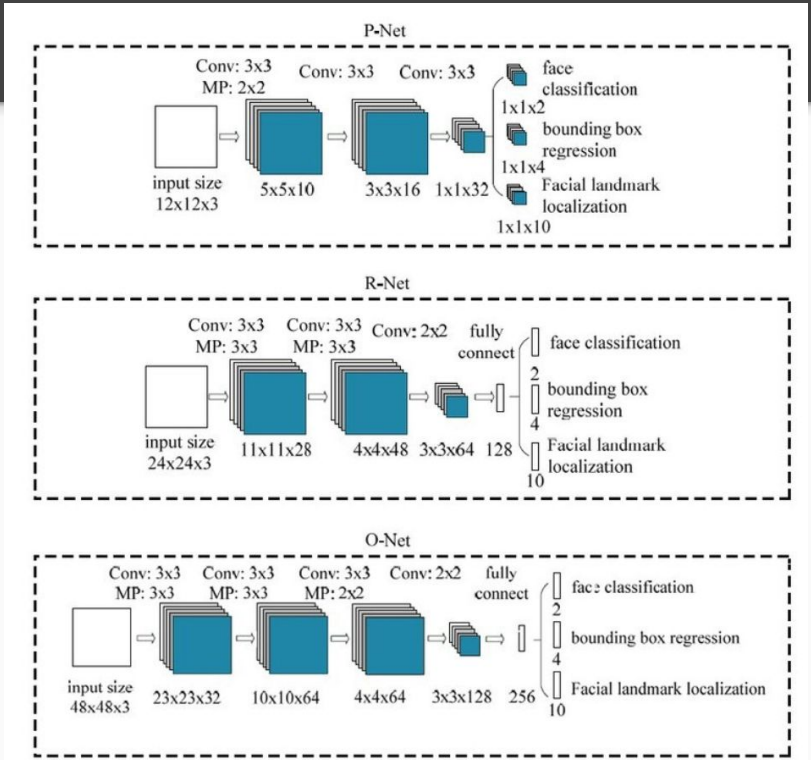


https://datascientest.com/quest-ce-que-le-modele-vgg



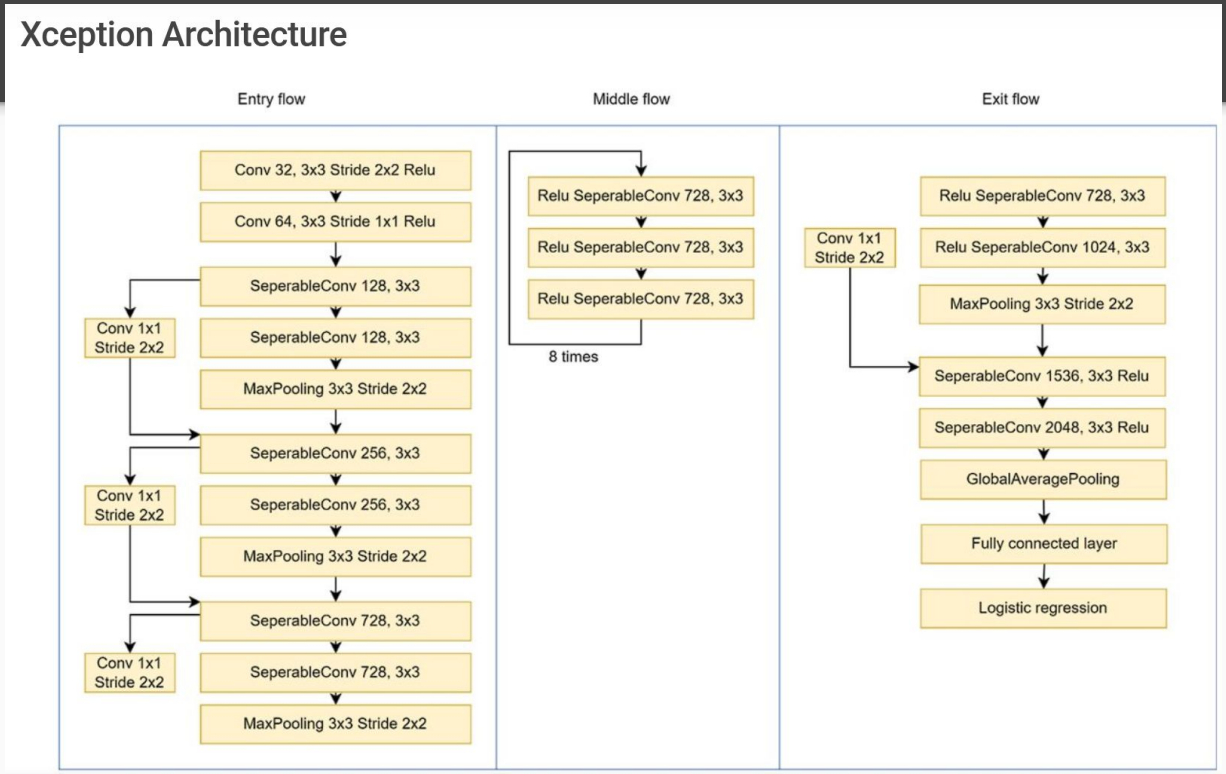
https://www.ikomia.ai/blog/mastering-resnet-deep-learning-image-recognition

Détection de deepfake - metric learning (MTCNN)



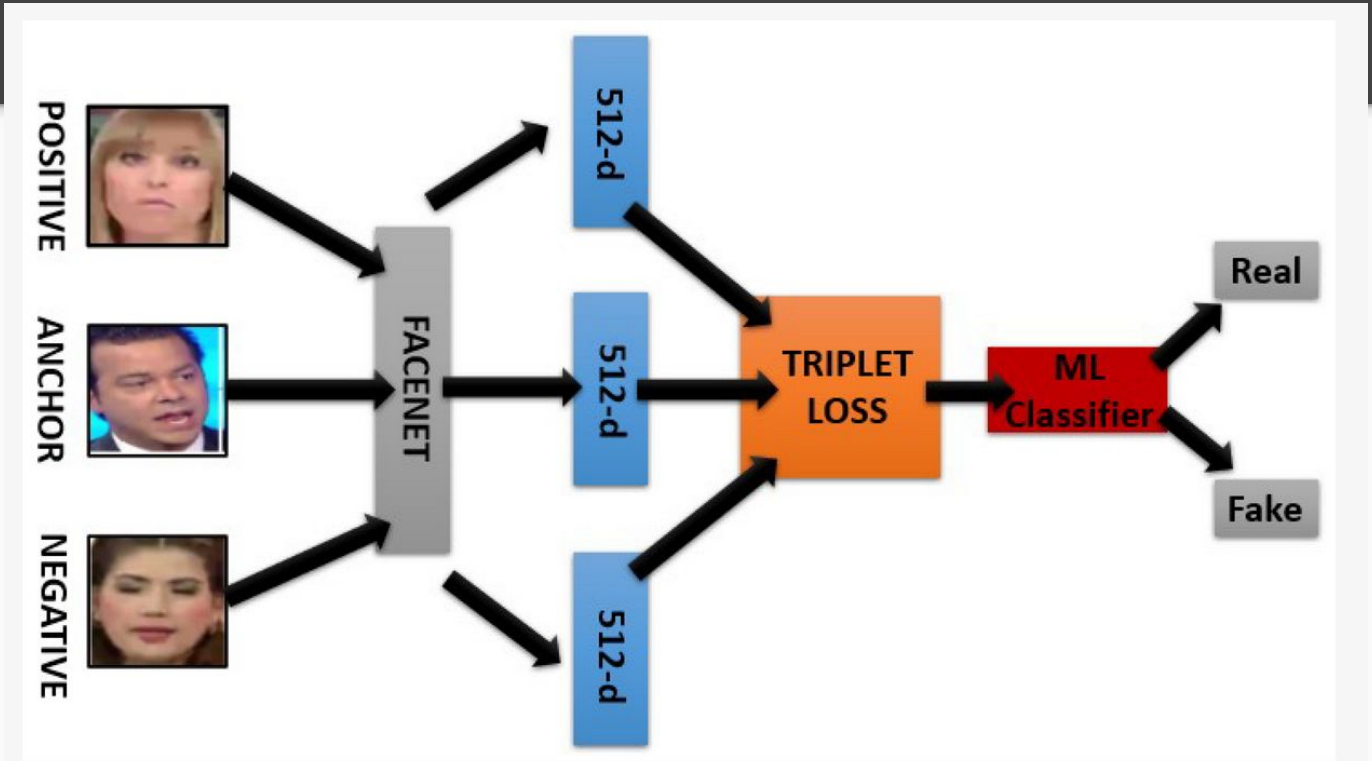
https://www.researchgate.net/figure/MTCNN-Stage-architecture-of-the-model-used-for-face-detection-and-landmark-extraction_fig3_341148320

Détection de deepfake - metric learning (Xception)



https://link.springer.com/content/pdf/10.1007/s10479-022-05151-y.pdf

Détection de deepfake - metric learning (Apprentissage de Triplet)



Solutions pour prévenir les abus

- Donner les moyens de comprendre (sensibilisation, expérimentation, Ressources accessibles)
- Responsabiliser (Bonnes pratiques, IA interprétables)
- Observer (usages, dérives..)

Lien avec la cognition

- Reproduction des biais cognitifs : apparition des biais

- Auto-amélioration : Fonctionnement des GAN qui permet d'améliorer ses productions sans supervision

- Embodiment : Reproduction de visage/voix réaliste pour simuler l'interaction humaine

Question ?

Quelle est l’image généré par l’IA

Quelle est l’image généré par l’IA

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

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

Les États ? Les entreprises à qui appartiennent les IA ? Les deux ?

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

Influence sur la société ? Risques potentiels ?

Télé réalité coréenne utilise l’IA pour ramener cette fille à la vie

Sources

https://linc.cnil.fr/le-tatouage-numerique-une-mesure-de-transparence-salutaire-22

https://info.haas-avocats.com/droit-digital/la-lutte-contre-les-deepfakes-sorganise

https://www.mdpi.com/2079-9292/13/1/95

https://www.minalogic.com/livre-blanc-ia-generative-hypertrucages-deepfake/

https://www.cnetfrance.fr/news/ai-ou-reel-devinez-si-ce-sont-de-vraies-photos-394872.htm

https://link.springer.com/content/pdf/10.1007/s10479-022-05151-y.pdf