La Statistique et l'IA générative

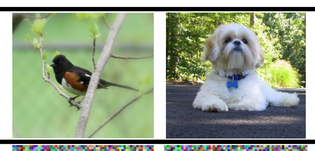
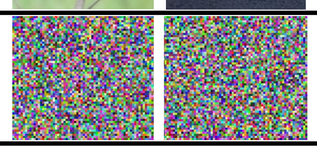
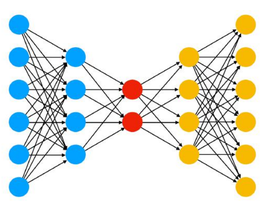
Les modèles génératifs sont aujourd'hui omniprésents :
pour traduire, pour compléter des informations manquantes, ou pour créer, ils sont utilisés dans la sphère publique et privée.
Mais comment fonctionnent-ils ?

COMMENT L'IA APPREND À GÉNÉRER ?

Une IA générative correspond à ce qu'on appelle un **modèle génératif profond**, un cas particulier de modèle **probabiliste**. Attardons-nous un peu sur ces termes…
- **Génératif** car il est capable de **générer des nouvelles données**.
- **Profond** car il utilise comme brique élémentaire des **réseaux de neurones profonds**, des fonctions mathématiques inspirées des réseaux de neurones biologiques.
- **Probabiliste** car sa manière de **gérer l'incertitude** (et donc de générer des données variées) fait appel à la théorie des probabilités.

Afin que notre modèle soit capable de générer des données réalistes, nous lui montrons une grande base de données réelles (appelée données d'entrainement). Les connections de ses réseaux de neurones (les paramètres du modèle) sont alors optimisés. C'est ce qu'on appelle la **phase d'entraînement du modèle**.

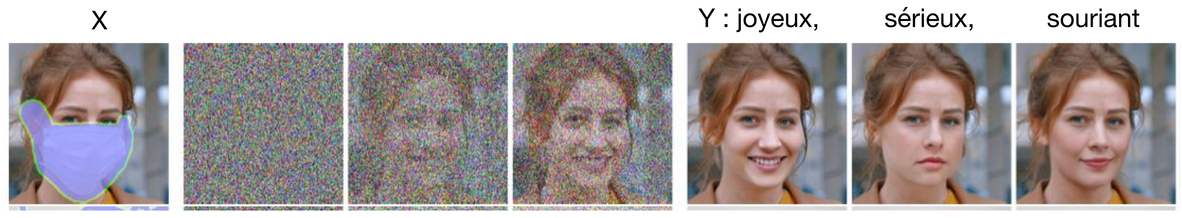
Comment cette phase d'entrainement fonctionne ?
Chaque modèle probabiliste est associé à une mesure de surprise, qui mesure à quel point les données sont surprenantes.
Pour les modèles de langage, la mesure de surprise est « à quel point le mot suivant est surprenant ? » ; pour les images, le modèle est capable de quantifier la surprise de n'importe quelle image (potentiellement incomplète).

https://www.image-net.org	Avant entrainement	Après entrainement
	**Très suprenant !**	**Pas suprenant**
	**Moins suprenant**	**Plutôt suprenant**
	Les connections neuronales sont choisies aléatoirement. Le modèle n'est pas performant sur les vraies images, et trouve toute donnée structurée très surprenante.	Le modèle reconnait les données qu'on lui a montrées, qui ne le surprennent plus. Il peut créer de nouvelles données réalistes, proches de ce qu'il a appris.

On considère ici des modèles non supervisés, où on observe seulement X et on modélise $p(x)$.
On peut avoir plus d'informations : une classe ou une caractéristique Y associée à X que l'on connait sur les données utilisées pendant la phase d'entraînement du modèle. Alors, on cherche à modéliser la loi conditionnelle $p(y \mid x)$.
Dans les deux cas, l'IA va estimer ces lois : $\hat{p}(x)$ et $\hat{p}(y \mid x)$.

COMMENT RECONSTRUIRE UN VISAGE ?

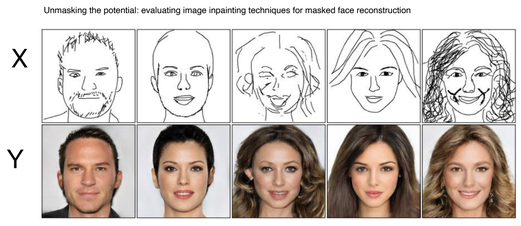
Lorsqu'il est masqué
On veut savoir si une personne sourit derrière son masque. Pour cela, on va utiliser toute la partie visible du visage: on dit que certaines personnes sourient avec les yeux ! Le lien entre le haut et le bas du visage peut être appris par un modèle génératif à partir d'un grand nombre d'exemples, comme on le fait intuitivement en voyant des visages humains depuis notre naissance.



RePaint: Inpainting using Denoising Diffusion Probabilistic Models

En mathématiques, une image est représentée par un tableau comprenant autant de valeurs que de pixels. Le masque supprime une partie de ces données : on a donc des valeurs manquantes. Le but est de les remplacer par des valeurs plausibles en utilisant les valeurs que l'on observe dans le tableau, et donc d'apprendre la loi de probabilité des données manquantes sachant les données observées. C'est ce qu'on appelle **l'imputation**.

À partir d'un croquis
On peut aussi vouloir découvrir une photographie probable d'une personne dont on n'a qu'un croquis (qui est une version dégradée de la photo), ça peut servir à la police !
C'est un problème difficile : c'est pourquoi on doit mieux le poser, en donnant des règles a priori, par exemple : on sait que le visage humain a certaines caractéristiques (deux yeux, un nez, une bouche, ...).

Unmasking the potential: evaluating image inpainting techniques for masked face reconstruction


En mathématiques, la règle a priori est une loi de probabilité sur les visages. On va la combiner avec une vraisemblance pour obtenir une loi de probabilité, dite a posteriori, dépendant du croquis de l'image que l'on souhaite restaurer. Ce sont ce qu'on appelle des **problèmes inverses**. Des travaux proposent aussi de quantifier l'incertitude de l'image reconstruite à partir de la loi a posteriori, ce qui peut être utile dans notre exemple pour ne pas se diriger vers le mauvais suspect !

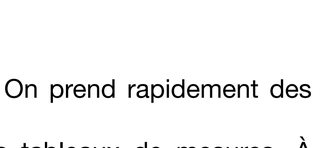
COMMENT L'IA GÉNÉRATIVE EST UTILISÉE EN MÉDECINE ?

Aux urgences
Quand un patient arrive dans un état critique, il faut agir vite. On prend rapidement des mesures, mais certaines sont parfois manquantes.
L'imputation est utilisée pour compléter automatiquement ces tableaux de mesures. À partir du tableau complété, il est plus facile de répondre à certaines questions, comme la réaction par rapport à un traitement.

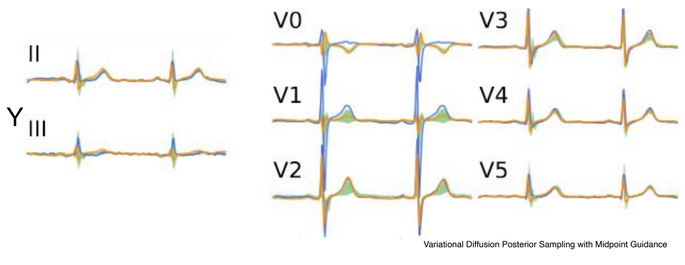
Reconstruction IRM
En IRM ou en scanner, plus l'examen est court, mieux c'est pour le patient. Mais si on prend moins d'images, elles sont souvent incomplètes ou floues.
Les problèmes inverses sont utilisés en imagerie médicale pour reconstruire une image claire à partir de peu de données. Cela permet de réduire les risques, notamment dûs à l'exposition aux rayons X.

Électrocardiogrammes sur montres connectées
Un des instruments clés pour les cardiologues est l'électrocardiogramme (ECG). Le cœur peut être vu comme une pompe, dont le but est d'envoyer le sang pour que toutes les cellules du corps humain aient de l'oxygène. Cette pompe fonctionne grâce à la propagation d'une impulsion électrique. L'ECG permet au médecin d'observer, à partir des électrodes mis sur la peau d'un patient, la propagation de cette onde. Un électrocardiogramme standard est constitué de 9 électrodes.

Des électrocardiogrammes peuvent aussi être réalisés avec certains modèles de montres connectées. Dans ce cas, seulement une électrode est utilisée, et on obtient un signal comme dans la figure ci-contre.



Des modèles génératifs peuvent reconstruire un ECG complet à partir des données de la montre, comme les signaux ci-après (en bleu le vrai signal, en orange un échantillon, en vert une bande de confiance).



Variational Diffusion Posterior Sampling with Midpoint Guidance

Utilisation des signaux reconstruits pour une analyse médicale (par exemple détection de crise cardiaque)

COMMENT CHATGPT ME RÉPOND-IL?

Les grands modèles de langage (LLMs), comme ChatGPT, utilisent des statistiques pour comprendre et générer du texte.
Leur objectif ? Prédire le mot suivant (Y) dans une phrase à partir des mots déjà écrits (X).
Avant, on utilisait des méthodes simples, comme les n-grams : le modèle regarde les 3 derniers mots, et choisit le mot qui a la probabilité la plus élevée d'arriver $\hat{P}(Y \mid X)$ d'après ce qu'il a vu dans l'entrainement.

Maintenant je suis …

Le modèle peut proposer *contente* ou *fatiguée*.

Plus on regarde des grands groupes de mots, plus cela demande de mémoire et de puissance de calcul.
ChatGPT utilise une structure de transformers, qui regarde tous les mots en même temps. La grande nouveauté a été l'attention, qui permet au modèle de se concentrer sur les parties importantes du contexte pour prédire le mot suivant.

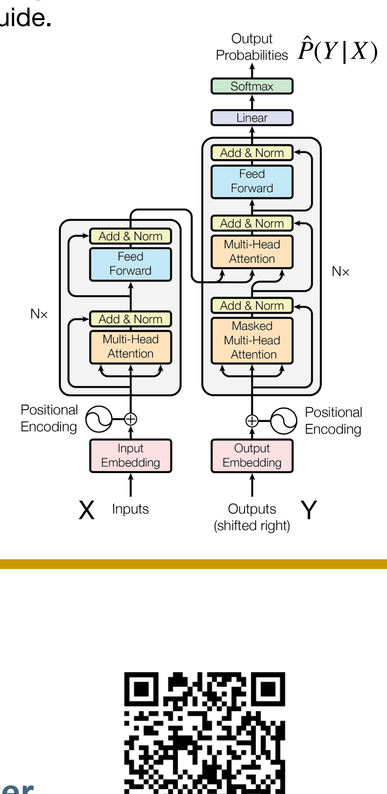
J'ai oublié mon parapluie ce matin, et maintenant je suis …

Le mot *parapluie* est utile pour deviner la suite : la personne est *mouillée* !

Dans X, l'entrée, on regarde potentiellement un long texte, pour prédire $\hat{P}(Y \mid X)$.
C'est cette architecture qui a permis de passer aux modèles géants comme ChatGPT, capables de lire, comprendre et générer du texte de manière fluide.

La structure de transformer (représentée à droite) est composée de plusieurs blocs, qui combinés permettent de détecter l'information importante.

Introduite en 2017, elle a marqué un tournant dans les modèles de traitement automatique de langue.
En rouge, les entrées / sorties du modèle.
En bleu, les neurones classiques.
En orange, le mécanisme d'attention.
En jaune, comment combiner les différents éléments.
En vert, la transformation pour obtenir une probabilité.



Attention Is All You Need

SFdS

La Société Française de Statistique assure la promotion de la statistique dans toutes ses composantes : RECHERCHE, ENSEIGNEMENT, APPLICATIONS.

Retrouvez sur notre site nos activités **: organisation d'évènements (conférences, soirées débats,..) , publications scientifiques, formation.**

Scannez moi pour télécharger ce poster