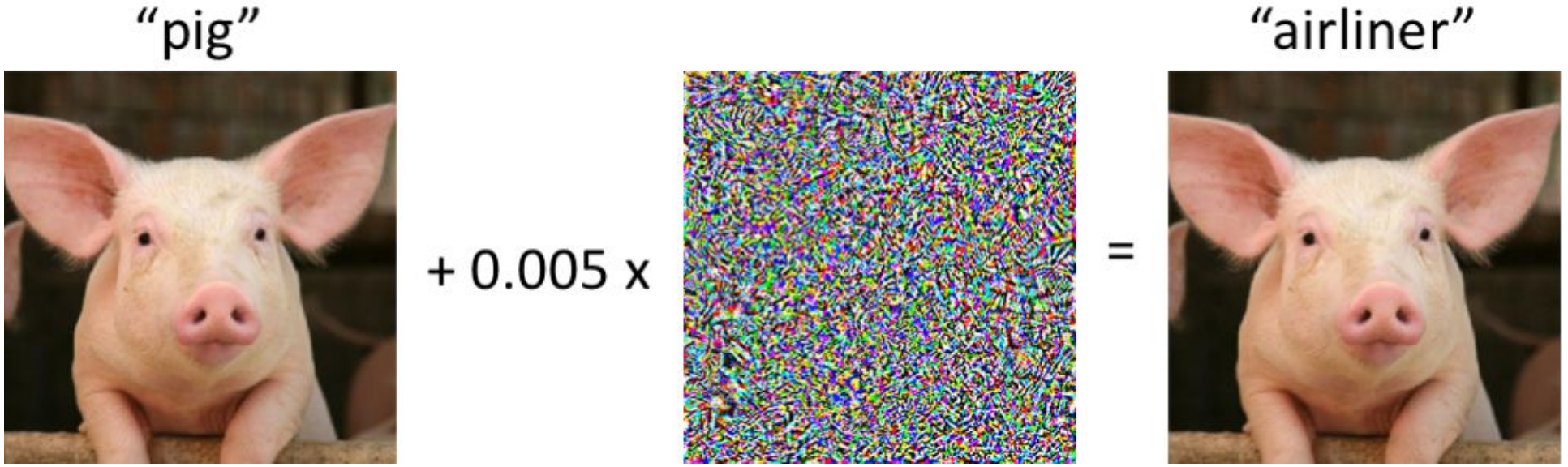
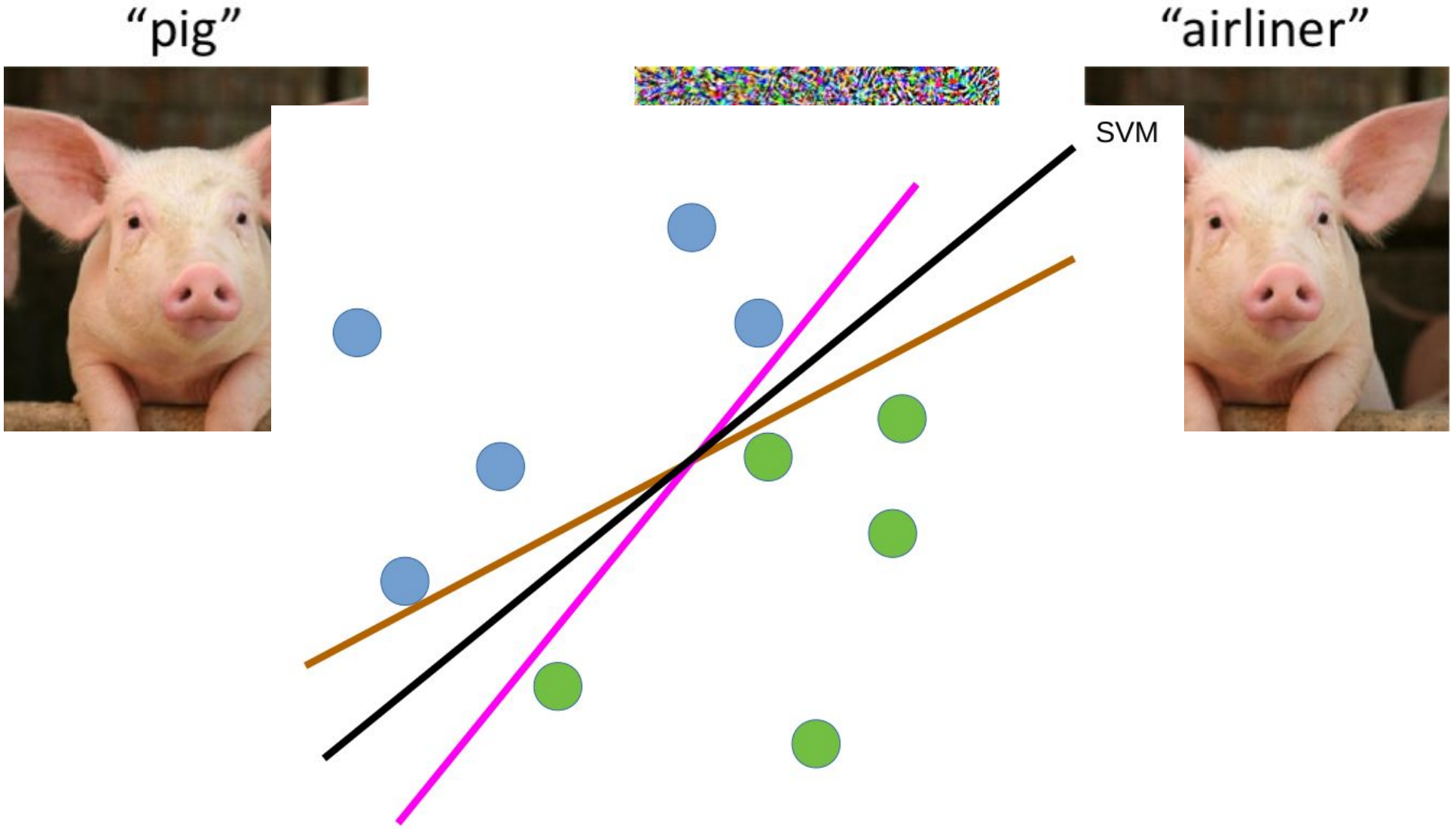
# Limites et perspectives

# Résistance face aux attaques

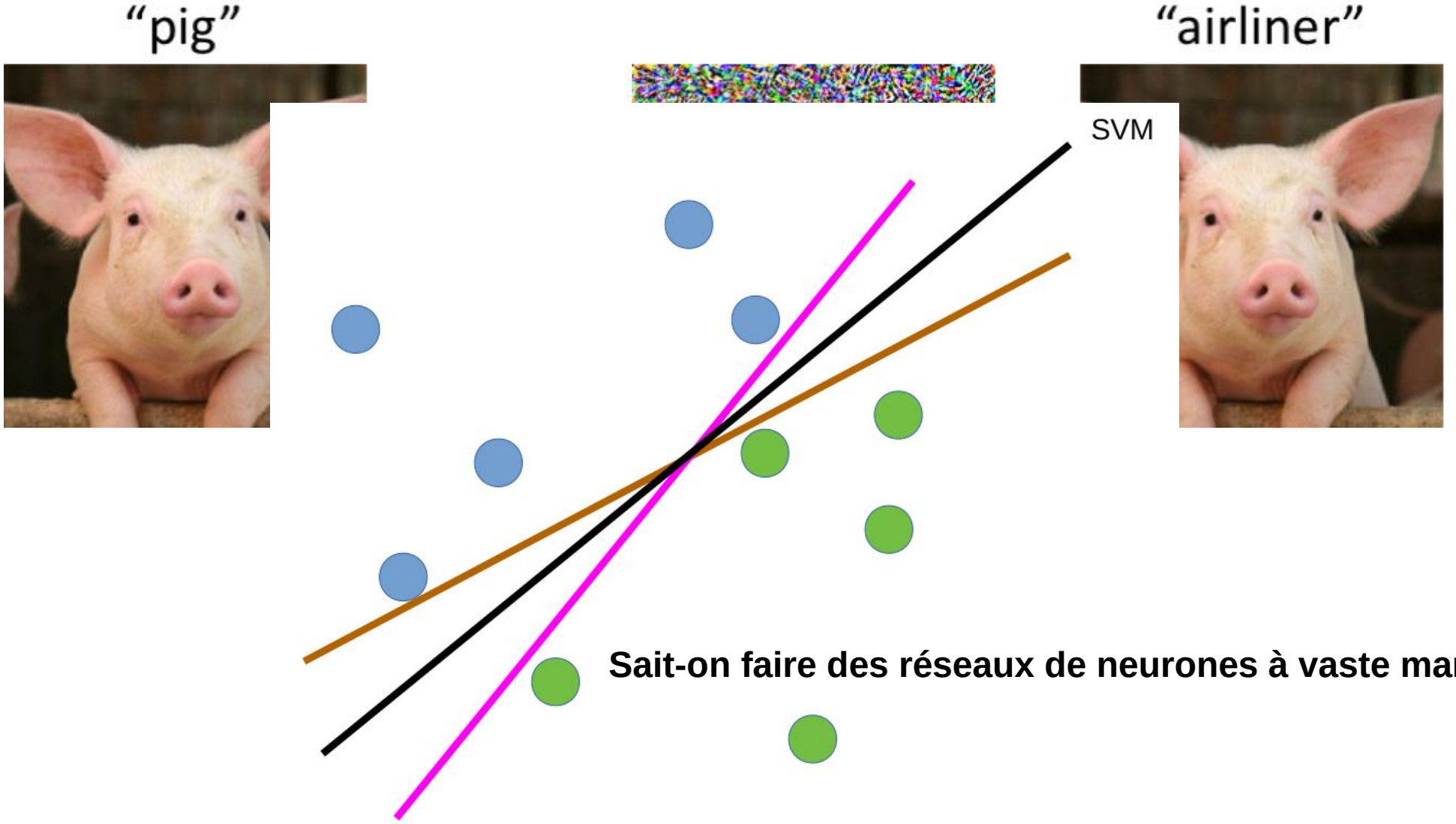



→ perte de performance critique

# Résistance face aux attaques

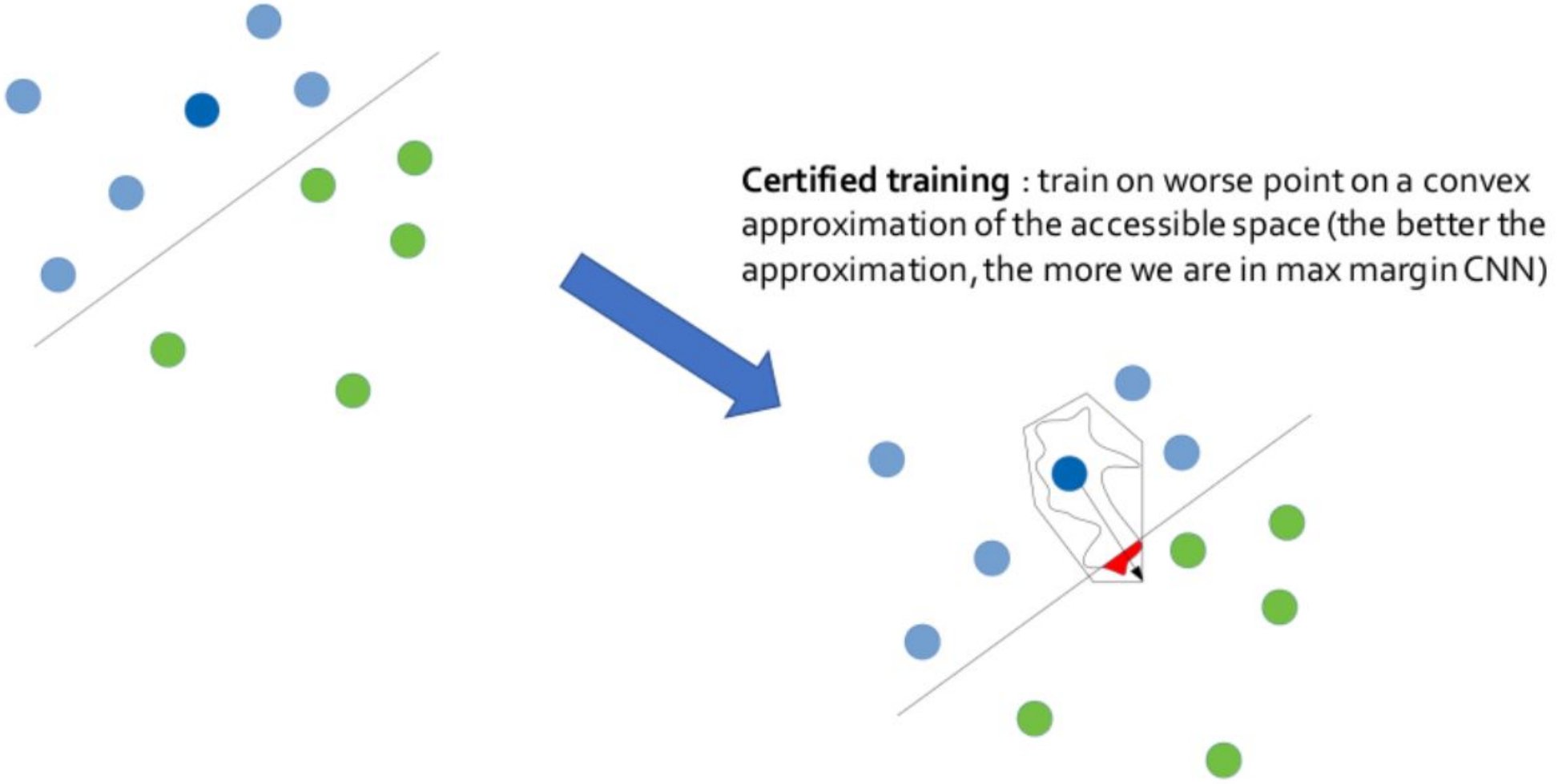


# Résistance face aux attaques



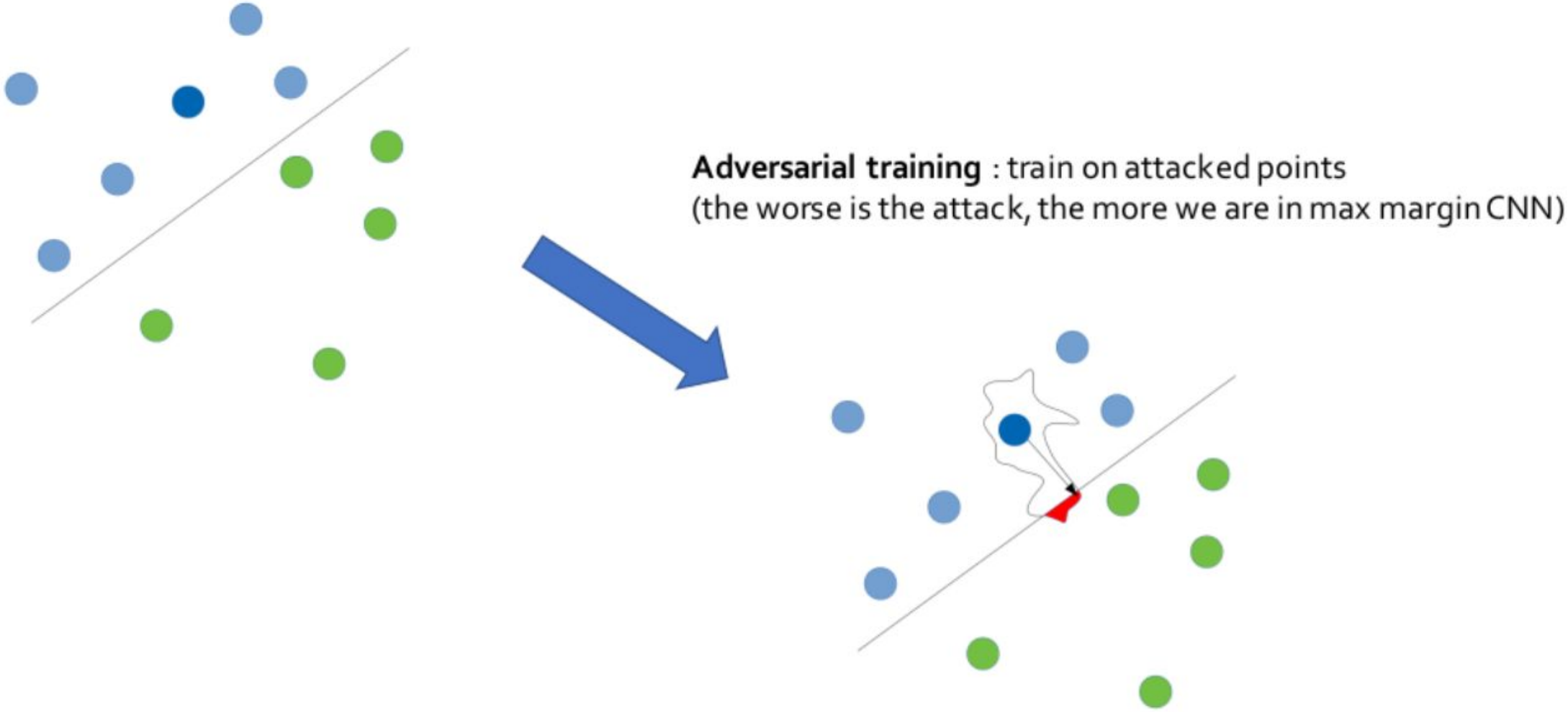

**Sait-on faire des réseaux de neurones à vaste marge ?**

# Résistance face aux attaques

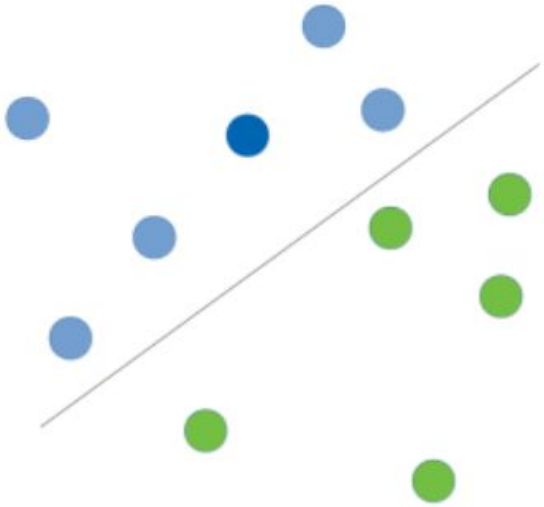



provable defenses against adversarial examples via the convex outer adversarial polytope

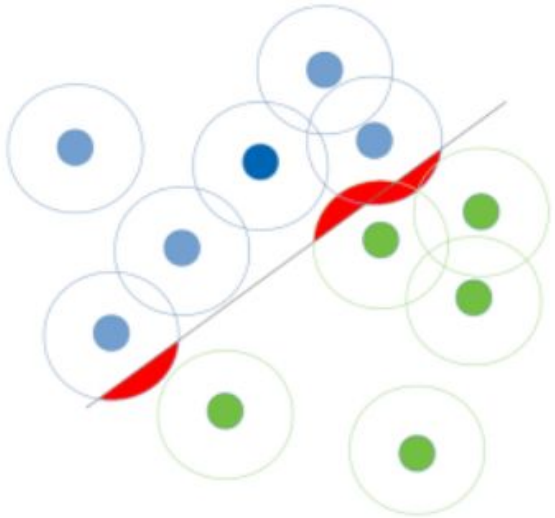
# Résistance face aux attaques



# Résistance face aux attaques



**Lipschitz training** :
a simple ball can be used to bound the accessible space !
(but requires specific architecture e.g. no relu)



Sorting Out Lipschitz Function Approximation

# Résistance face aux attaques

→ « best paper » NeurIPS 2021 : A universal law of robustness via isoperimetry

*Then, with probability at least* $1-\delta$ *with respect to the sampling of the data, one has simultaneously for all* $f \in \mathcal{F}$*:*

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 \leq \sigma^2 - \epsilon \;\Rightarrow\; \mathrm{Lip}(f) \geq \frac{\epsilon}{2^9\sqrt{c}}\sqrt{\frac{nd}{p\log(60WJ\epsilon^{-1})+\log(4/\delta)}}\,.$$

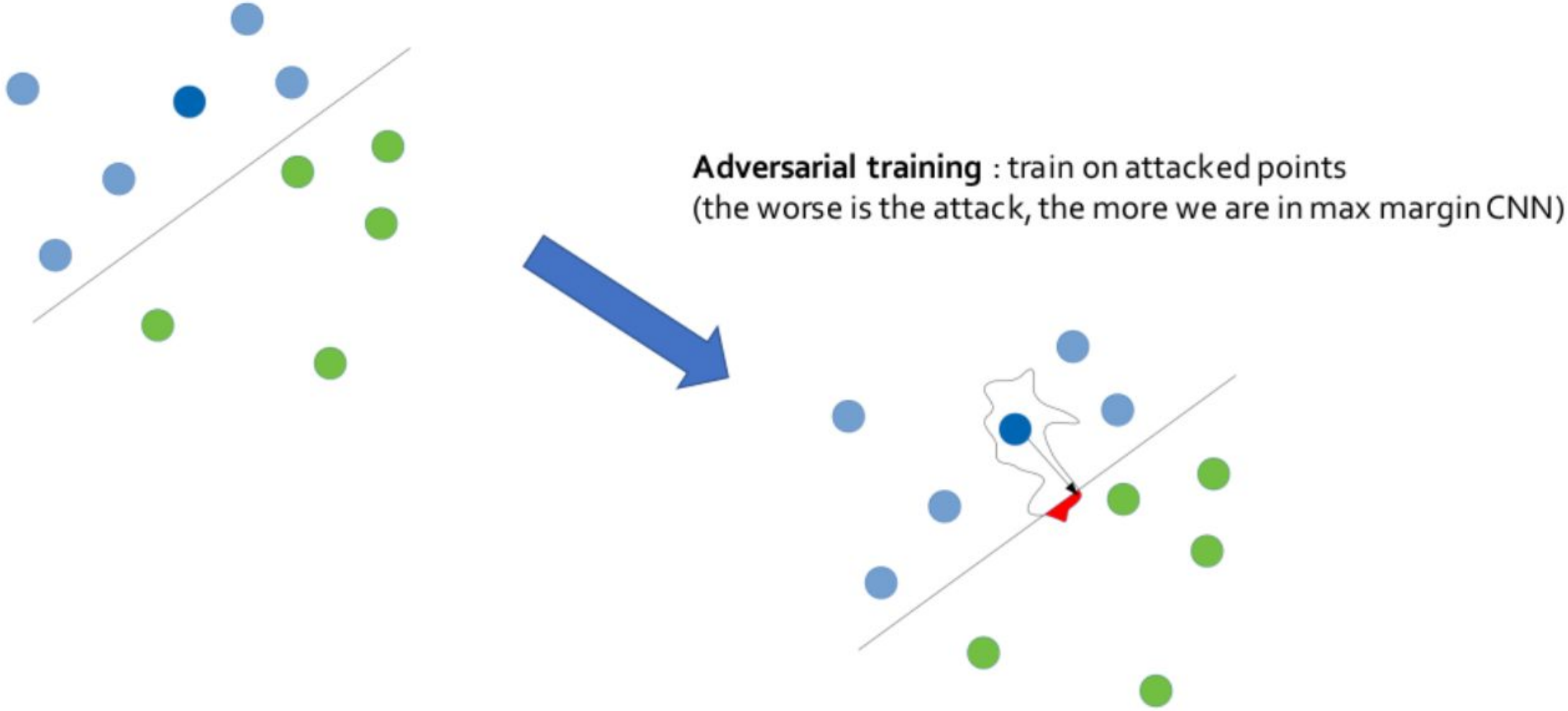
*Proof.* Define $\mathcal{W}_L \subseteq \mathcal{W}$ by $\mathcal{W}_L = \{\boldsymbol{w} \in \mathcal{W} : \mathrm{Lip}(f_{\boldsymbol{w}}) \leq L\}$. Denote $\mathcal{W}_{L,\epsilon}$ for an $\frac{\epsilon}{6J}$-net of $\mathcal{W}_L$. We have in particular $|\mathcal{W}_\epsilon| \leq (60WJ\epsilon^{-1})^p$. We apply Theorem 2 to $\mathcal{F}_{L,\epsilon} = \{f_{\boldsymbol{w}}, \boldsymbol{w} \in \mathcal{W}_{L,\epsilon}\}$:

$$\mathbb{P}\left(\exists f \in \mathcal{F}_{L,\epsilon} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \frac{\epsilon}{2} \text{ and } \mathrm{Lip}(f) \leq 2L\right)$$

$$\leq 4k\exp\left(-\frac{n\epsilon^2}{9^4 k}\right) + 2\exp\left(p\log(60WJ\epsilon^{-1}) - \frac{\epsilon^2 nd}{8^6 cL^2}\right)\,.$$

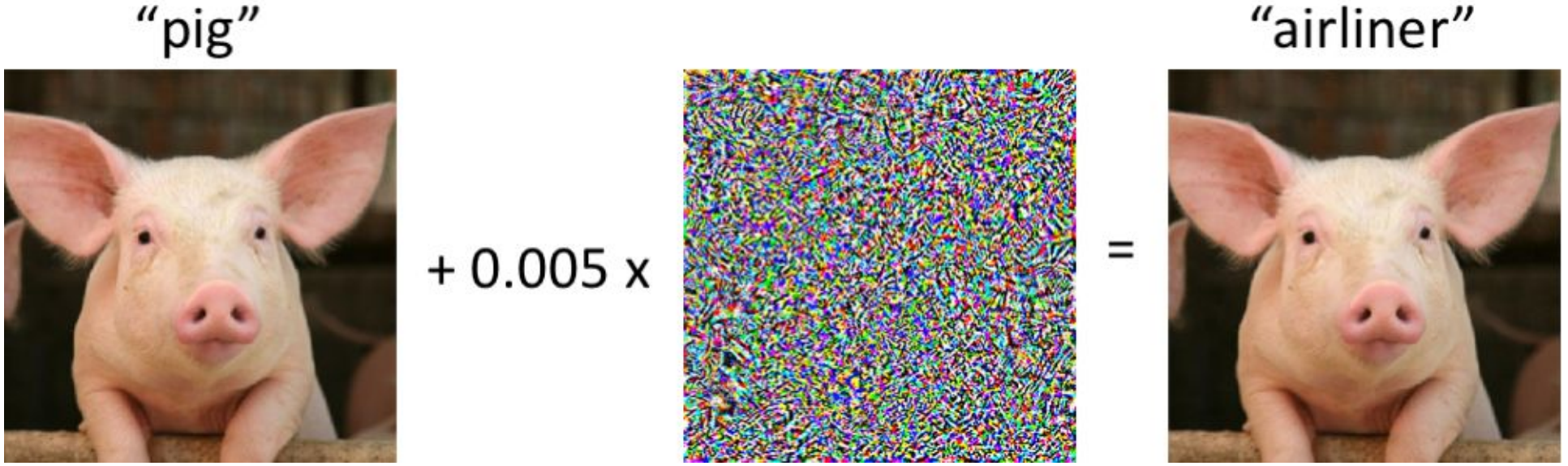
Observe that if $\|f-g\|_{\infty} \leq \frac{\epsilon}{4}$ and $\|y\|_{\infty}, \|f\|_{\infty}, \|g\|_{\infty} \leq 1$, then $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \frac{\epsilon}{2} +$

# Résistance face aux attaques



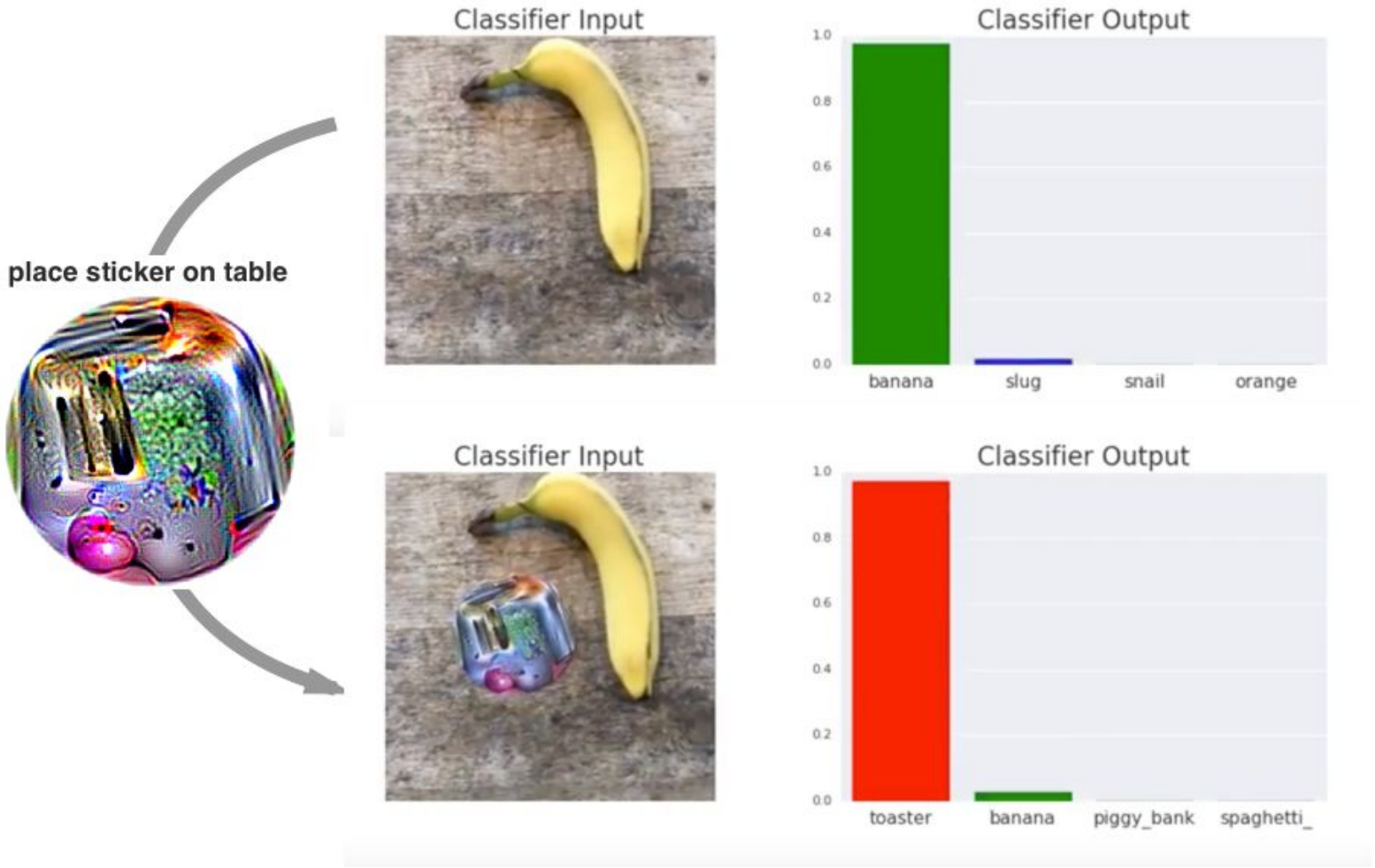

→ On sait créer des réseaux relativement « robustes » en pratique

# Attaques par patch



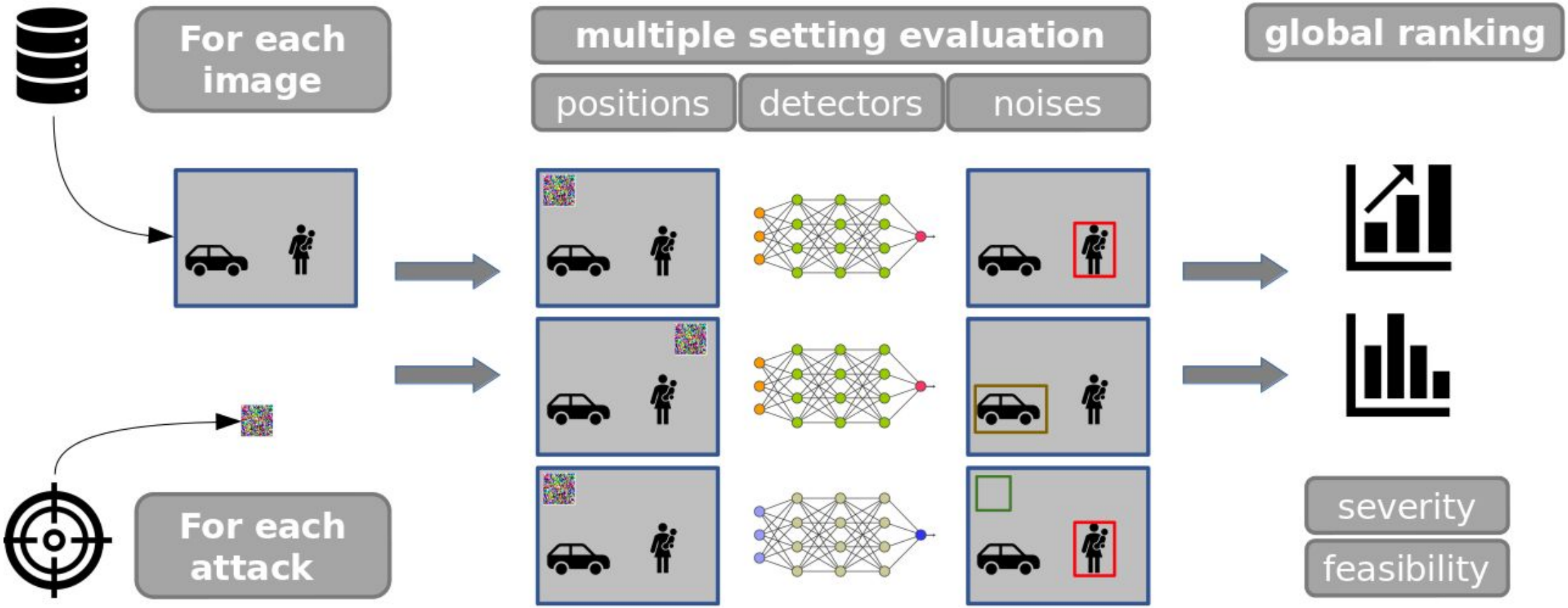

→ pas physiquement réaliste !

# Attaques par patch

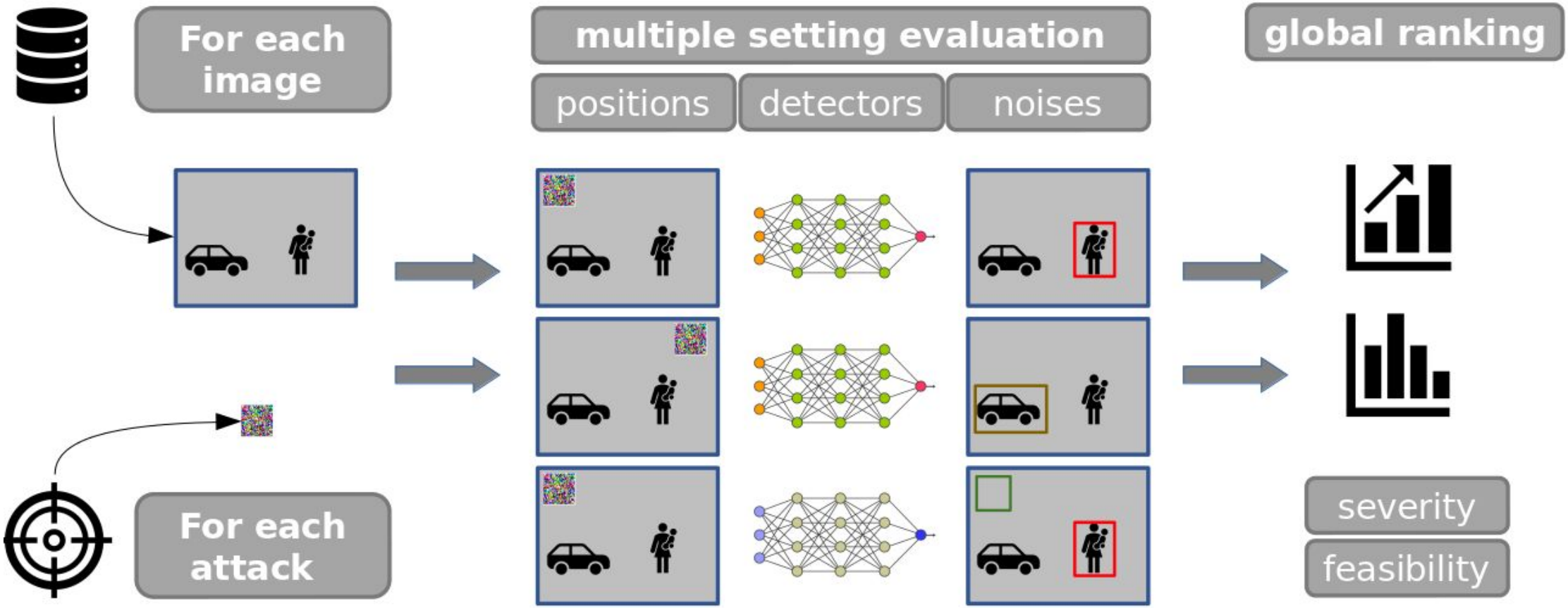


# Attaques par patch

# Attaques par patch

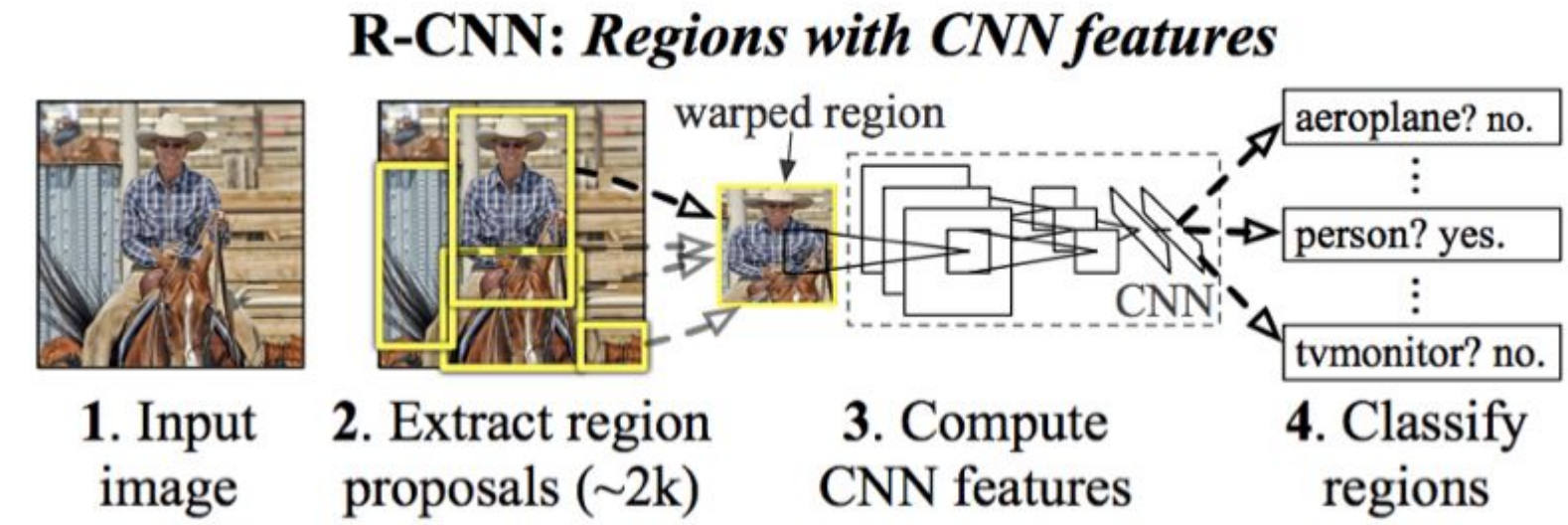


# Attaques par patch

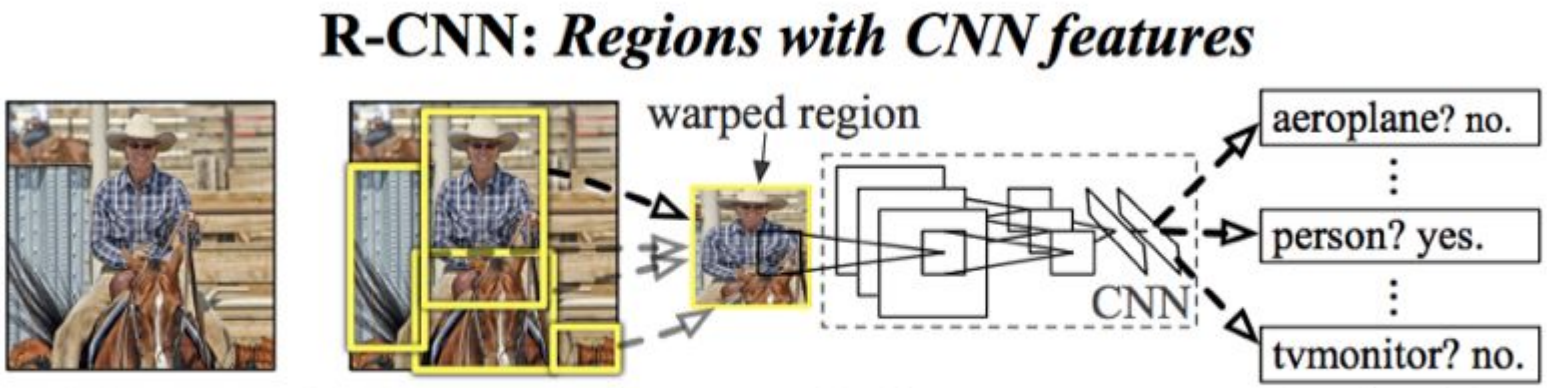



→ **il est possible de créer des attaques valides sous différentes positions !**

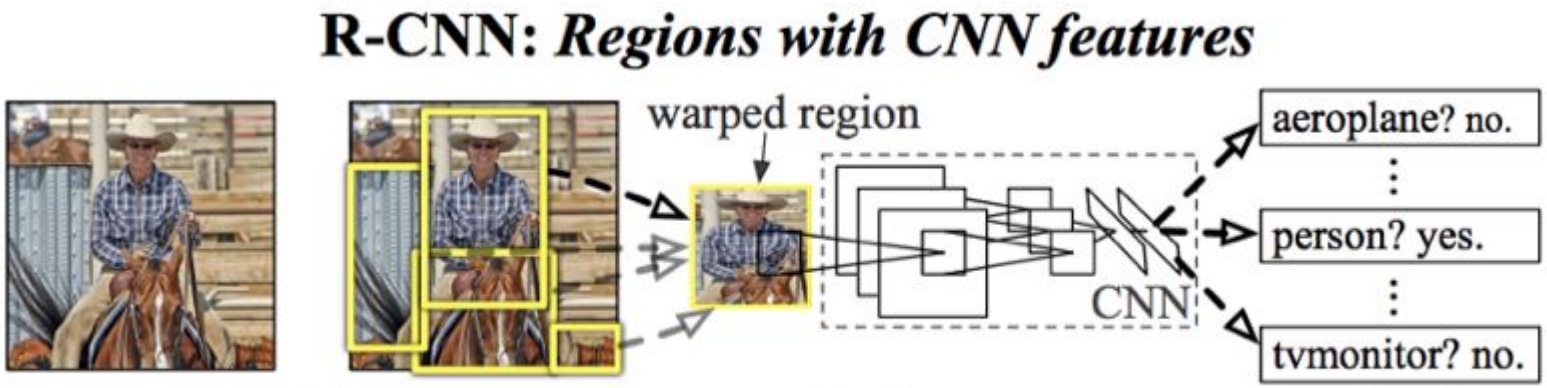
# Attaques par patch



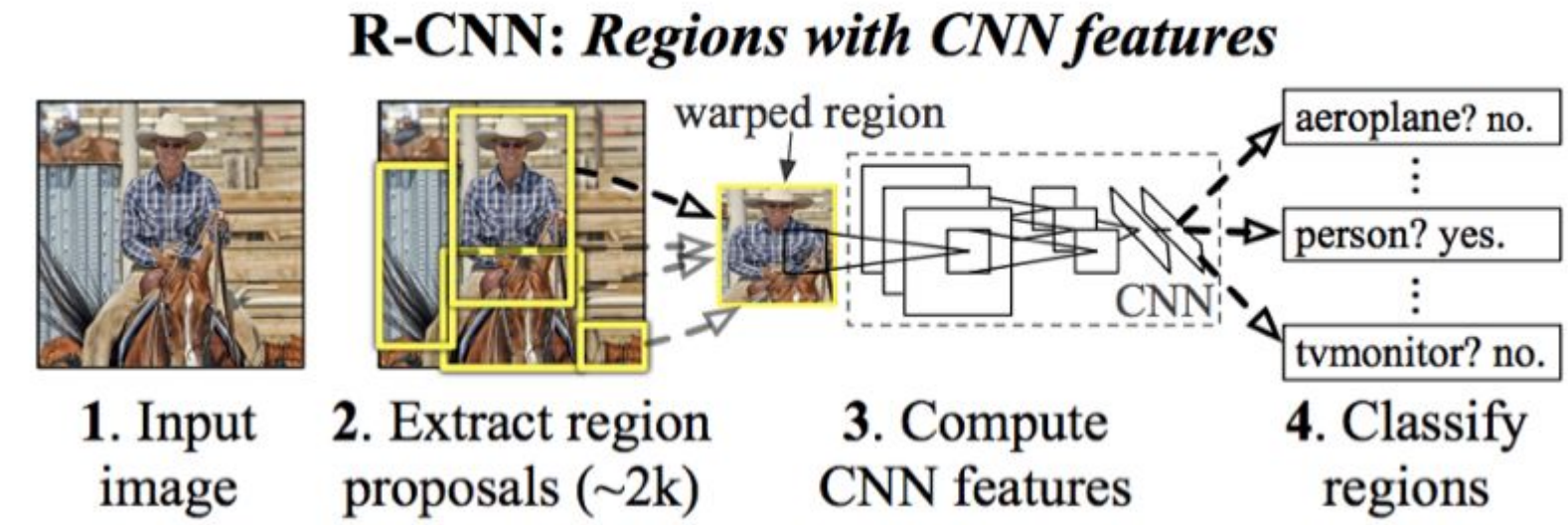
# Attaques par patch



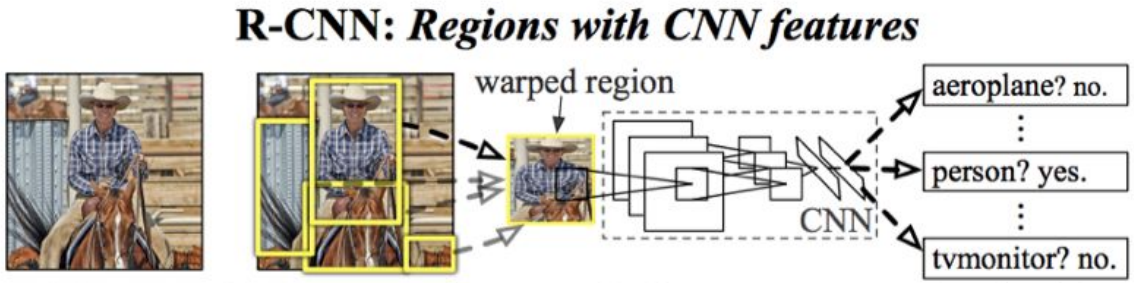

# Attaques par patch




# Attaques par patch



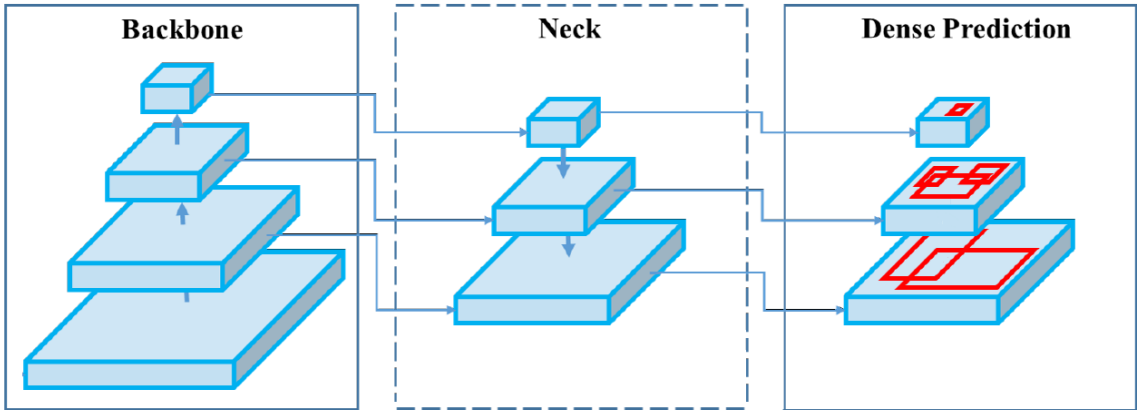

# Attaques par patch




Invariante aux attaques par patch.
Trop lente.

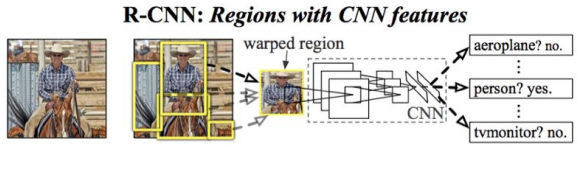
YOLO v4

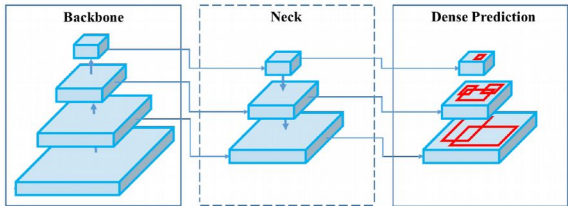
medium.com

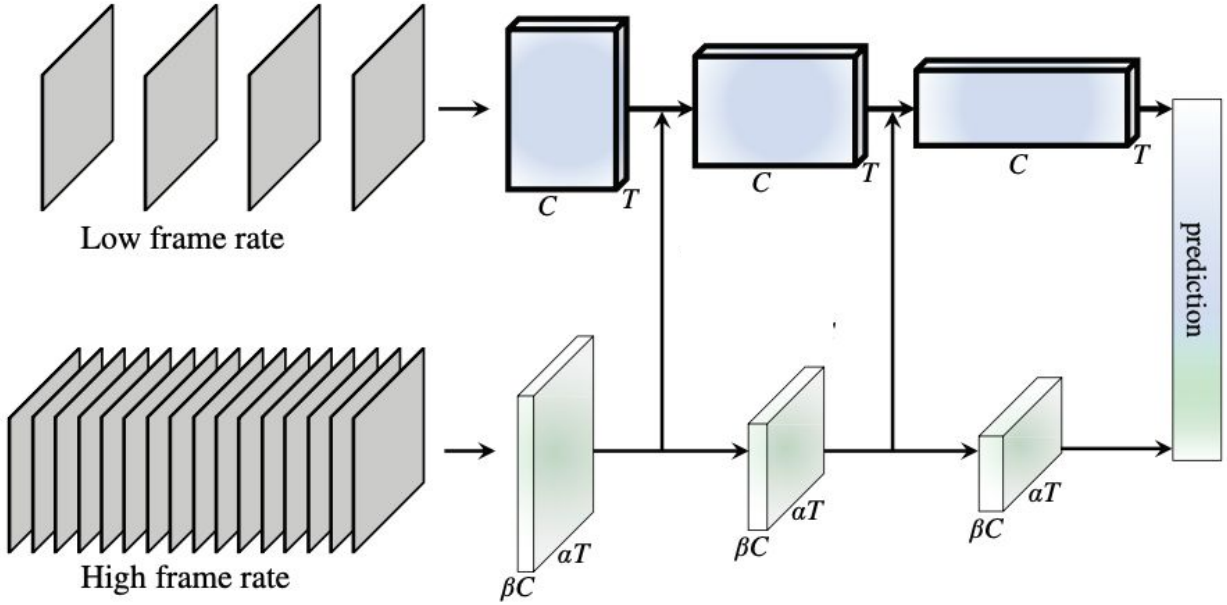



Sensible aux attaques.
Très rapide.

# Attaques par patch



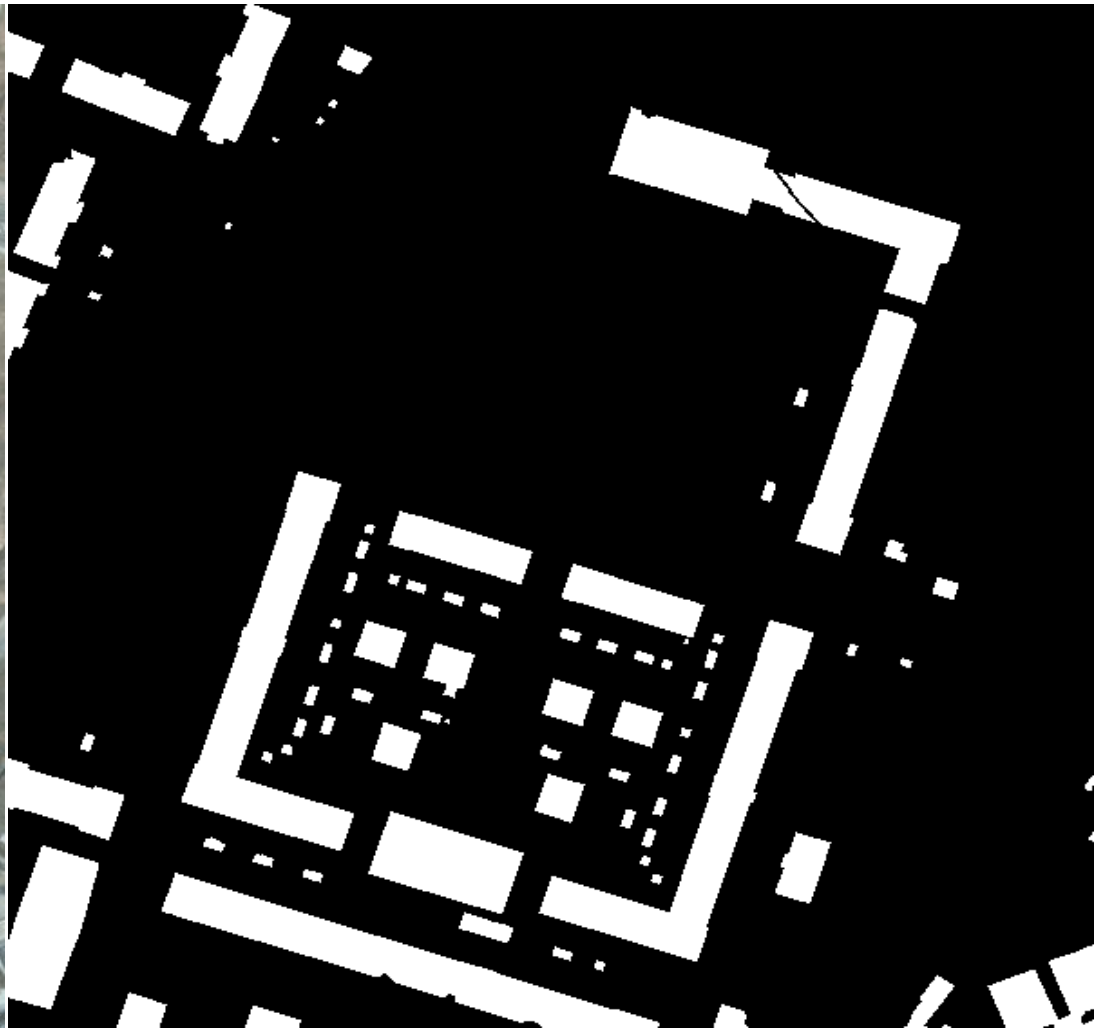



# Transfert learning

# Apprentissage par ordinateur : quelles garanties ?
# Transfert learning

ISPRS Potsdam dataset



# Apprentissage par ordinateur : quelles garanties ?
# Transfert learning

AIRS

# Apprentissage par ordinateur : quelles garanties ?
# Transfert learning

Deux images « similaires » pour nous…

# Apprentissage par ordinateur : quelles garanties ?
# Transfert learning

| Modèle (appris sur *AIRS train*) | Efficientnet-b7 |
|---|---|
| Performance sur *AIRS test* | 88 % |
| Performance sur *Potsdam* | 58 % |

# Apprentissage par ordinateur : quelles garanties ?
# Transfert learning

| Modèle (appris sur *AIRS train*) | Efficientnet-b7 | Histogrammes + 2 couches |
|---|---|---|
| Performance sur *AIRS test* | 88 % | 62 % |
| Performance sur *Potsdam* | 58 % | 56 % |

# Apprentissage par ordinateur : quelles garanties ?
# Données non IID

*Le vieux monde se meurt, le nouveau monde tarde à apparaître*
*et dans ce clair-obscur surgissent les monstres.*

*Antonio Gramsci*

# Apprentissage par ordinateur : quelles garanties ?
# Données non IID

*Le vieux monde se meurt, le nouveau monde tarde à apparaître*
*et dans ce clair-obscur surgissent les monstres.*

*Antonio Gramsci*

Étant donnée deux fonctions $f, y : X \to Y$ avec $X$ un espace de données muni d'une densité de probabilité $P$ et $Y$ un espace de label, et, $x_1, ..., x_K$, $K$ échantillons **tirés selon** $P$, alors, $\forall \delta \in ]0, 1[$

$$P\left(\int \mathbf{1}_{\neq}(f(x), y(x))P(x)\mathrm{d}x \leq \frac{\sum_k \mathbf{1}_{\neq}(f(x_k), y(x_k))}{K} + \sqrt{\frac{-\log(\delta)}{2K}}\right) \geq 1 - \delta$$

# Apprentissage par ordinateur : quelles garanties ?
# Données non IID

*Le vieux monde se meurt, le nouveau monde tarde à apparaître*
*et dans ce clair-obscur surgissent les monstres.*

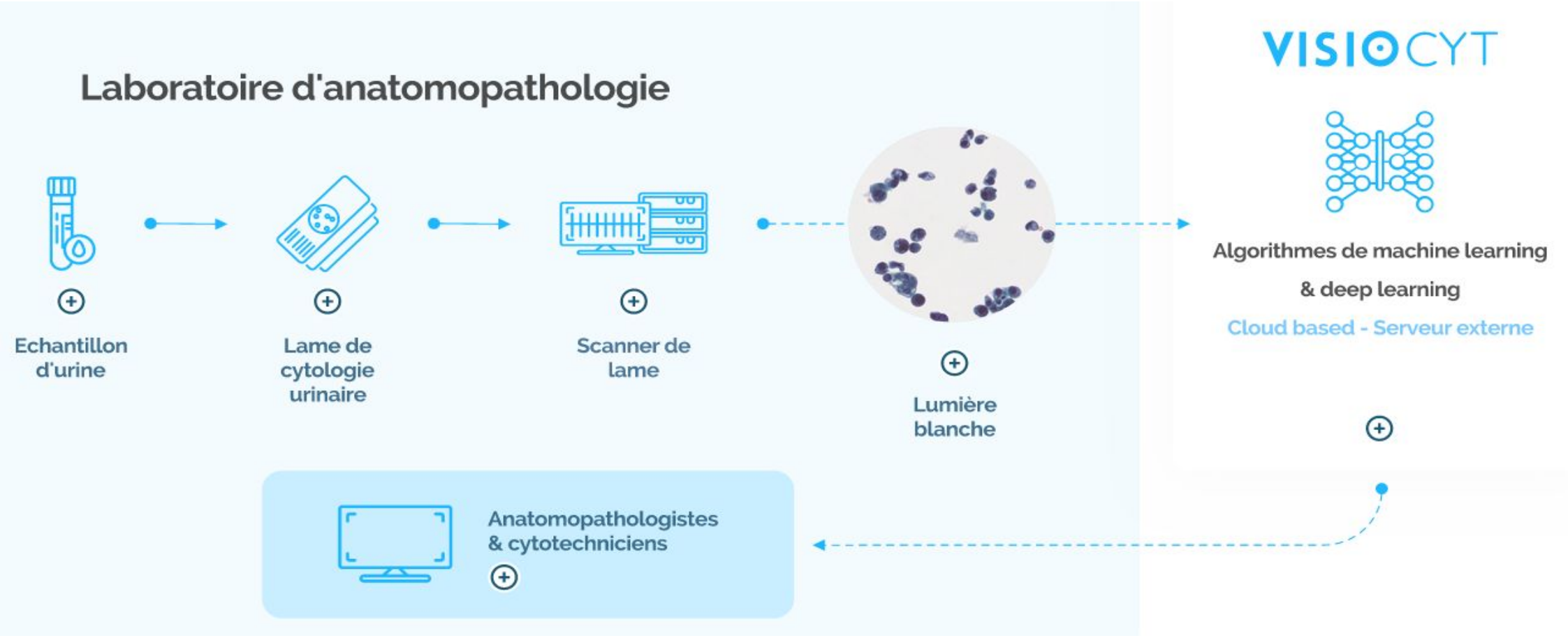
*Antonio Gramsci*

Étant donnée deux fonctions $f, y : X \to Y$ avec $X$ un espace de données muni d'une densité de probabilité $P$ et $Y$ un espace de label, et, $x_1, ..., x_K$, $K$ échantillons **tirés selon** $P$, alors, $\forall \delta \in ]0, 1[$

$$P\left(\int \mathbf{1}_{\neq}(f(x), y(x))P(x)\mathrm{d}x \leq \frac{\sum_k \mathbf{1}_{\neq}(f(x_k), y(x_k))}{K} + \sqrt{\frac{-\log(\delta)}{2K}}\right) \geq 1 - \delta$$

→ L'IA de confiance ?

# L’IA de confiance
# Les enjeux



# L’IA de confiance
# Les enjeux

**Aéronautique**

## ATTOL : des pilotes dans l’avion, mais un système autonome

**23/01/2020**

**L'ONERA a contribué à une 1ère mondiale en aéronautique civil : les premières campagnes d'essais de décollage autonome basé vision réalisées sur l'A350-1000 d'AIRBUS. On ne parle certes pas de drones, mais le système est bel et bien robuste et autonome.**

Pour le projet ATTOL (Autonomous Taxi, Take-Off and Landing), le 18 décembre dernier, une première mondiale a été accomplie : un décollage entièrement autonome basé vision, sans utilisation de l'ILS, ni du GPS, a été réalisé à plusieurs reprises sur l'aéroport de Toulouse-Blagnac. L'ONERA a contribué au développement et à la mise au point l'algorithme de fusion de données, qui élabore le signal de déviation à l'axe de piste, nécessaire au contrôle de l'avion, à

signaux se trouvent hors de l'appareil, ceux-c

A court terme également, des essais en vol d'
intégrant plusieurs développements de l'ONE
valider la solution algorithmique définie au c

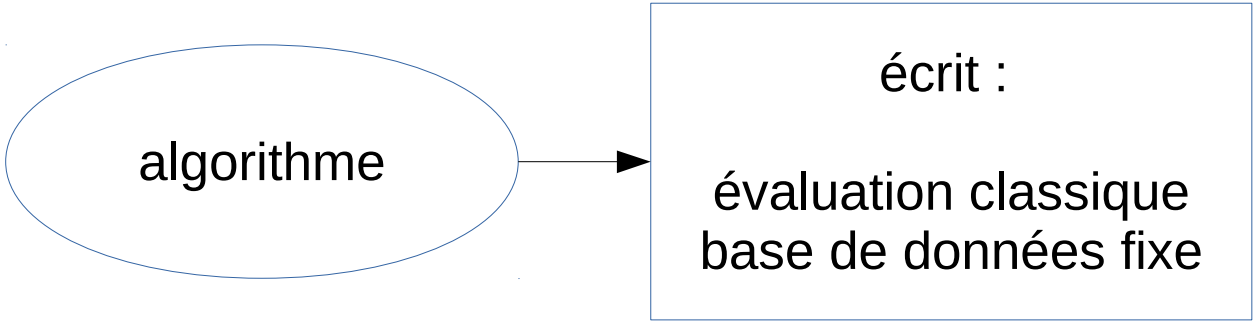
Une prouesse qui n’aurait pu être réalisée sa
dont certains sont détachés chez l’avionneur

# L’IA de confiance
# Les enjeux

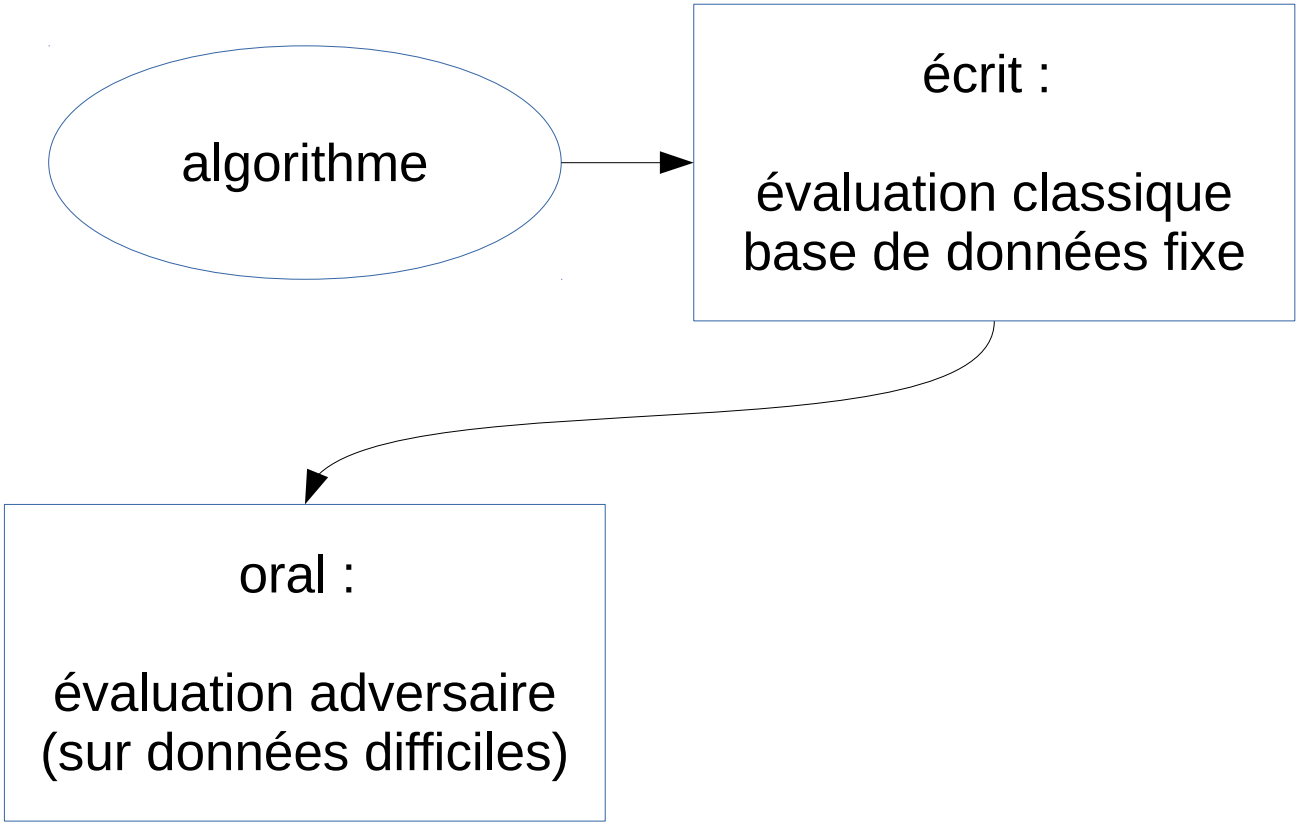
www.unmannedairspace.info

# Des pistes pour l'IA de confiance
# Évaluer les biais de tirage

# Des pistes pour l'IA de confiance
# Évaluer les biais de tirage

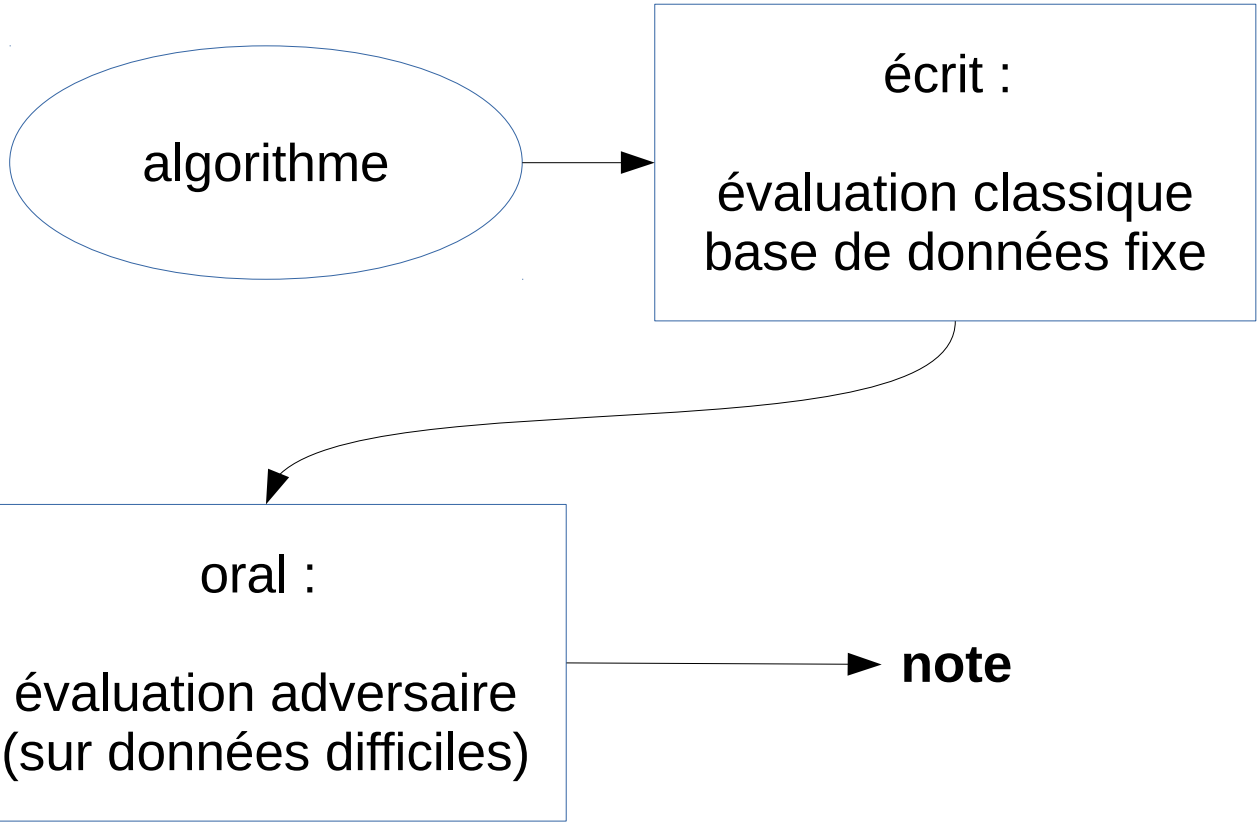



# Des pistes pour l'IA de confiance
# Évaluer les biais de tirage

# Des pistes pour l'IA de confiance
# Évaluer les biais de tirage

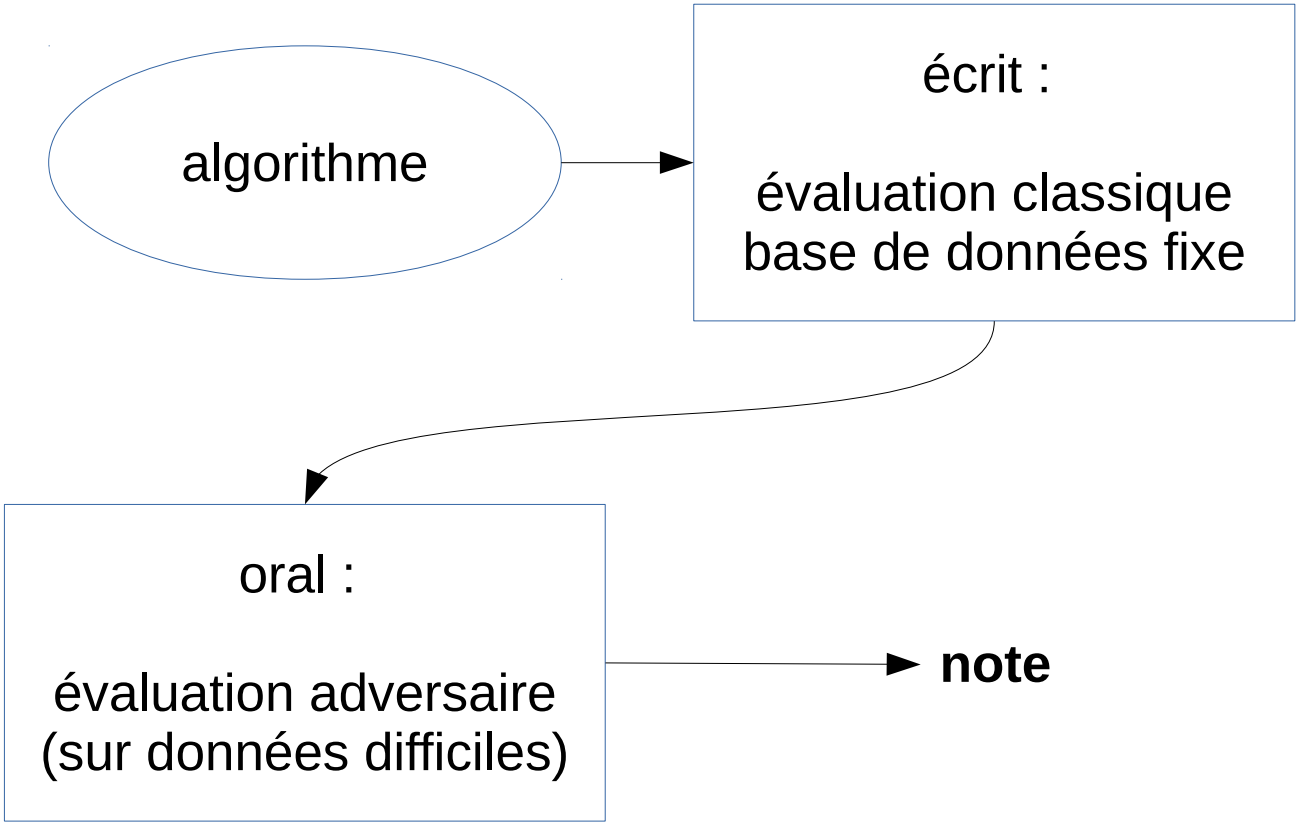
LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS



# Des pistes pour l'IA de confiance
# Évaluer les biais de tirage

LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS

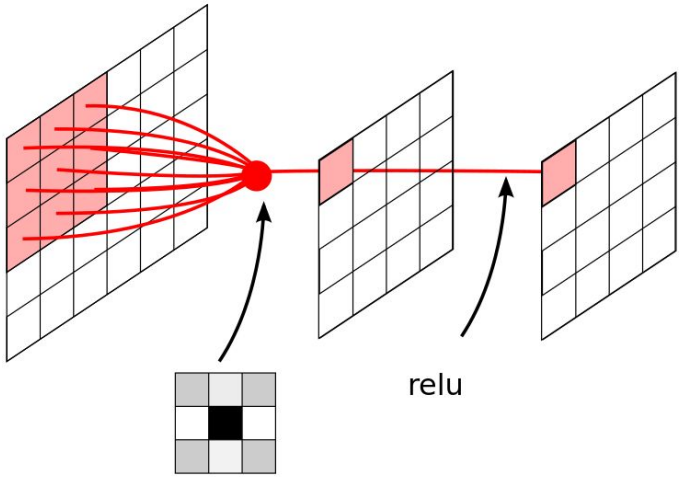
Comment vérifier que que cet « oral » mesure bien
la propension à être sensible à des données non IID ?

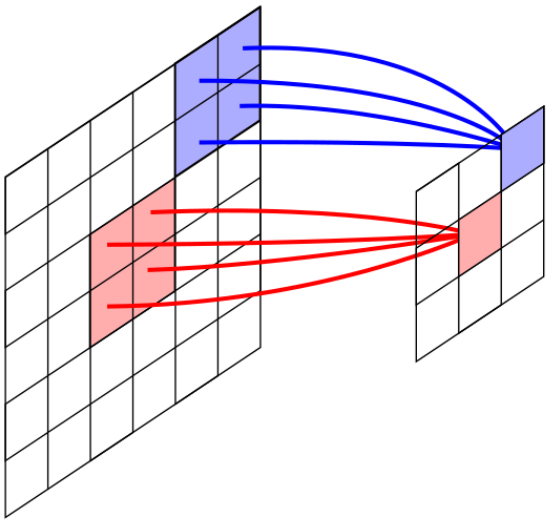


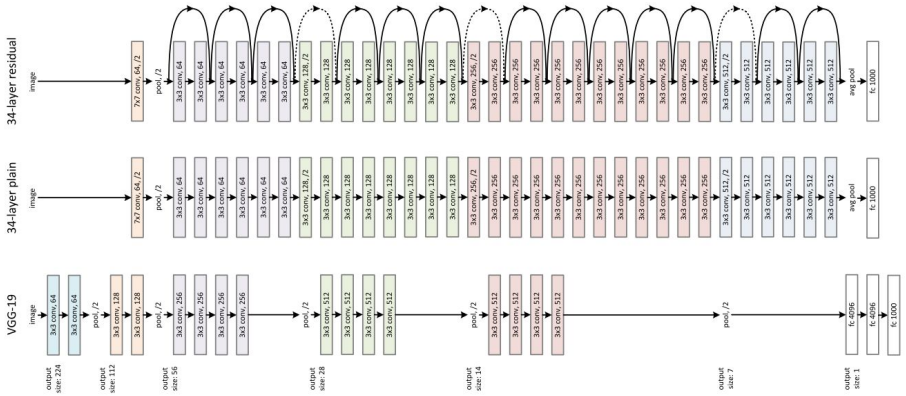
# Limites et perspectives

→ certification / IA de confiance

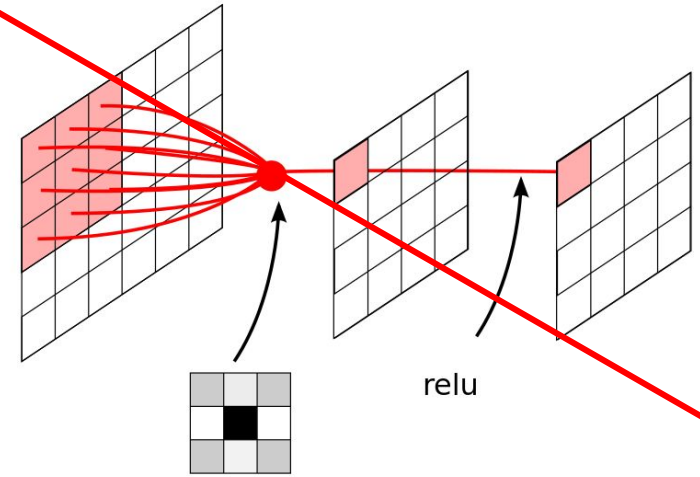
# Architecture transformer

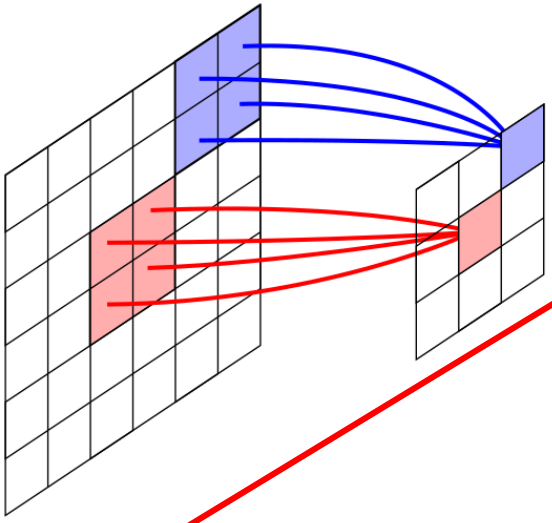


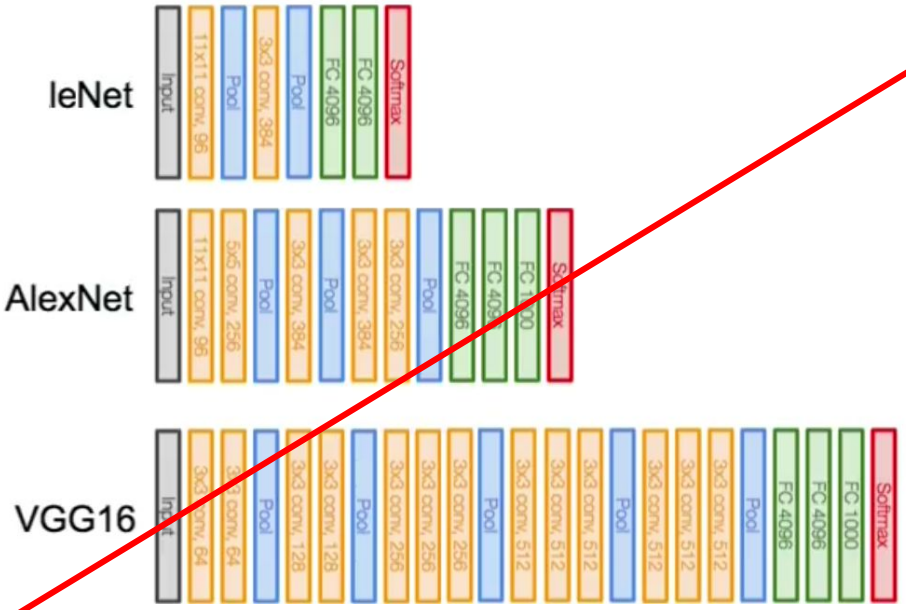



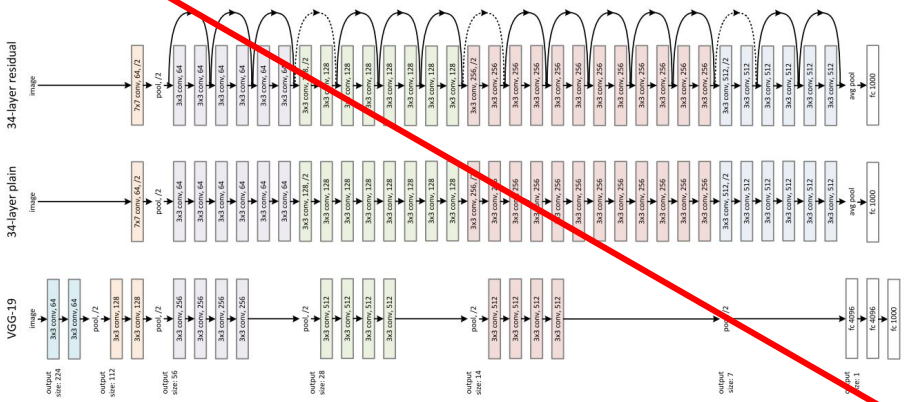



# Architecture transformer

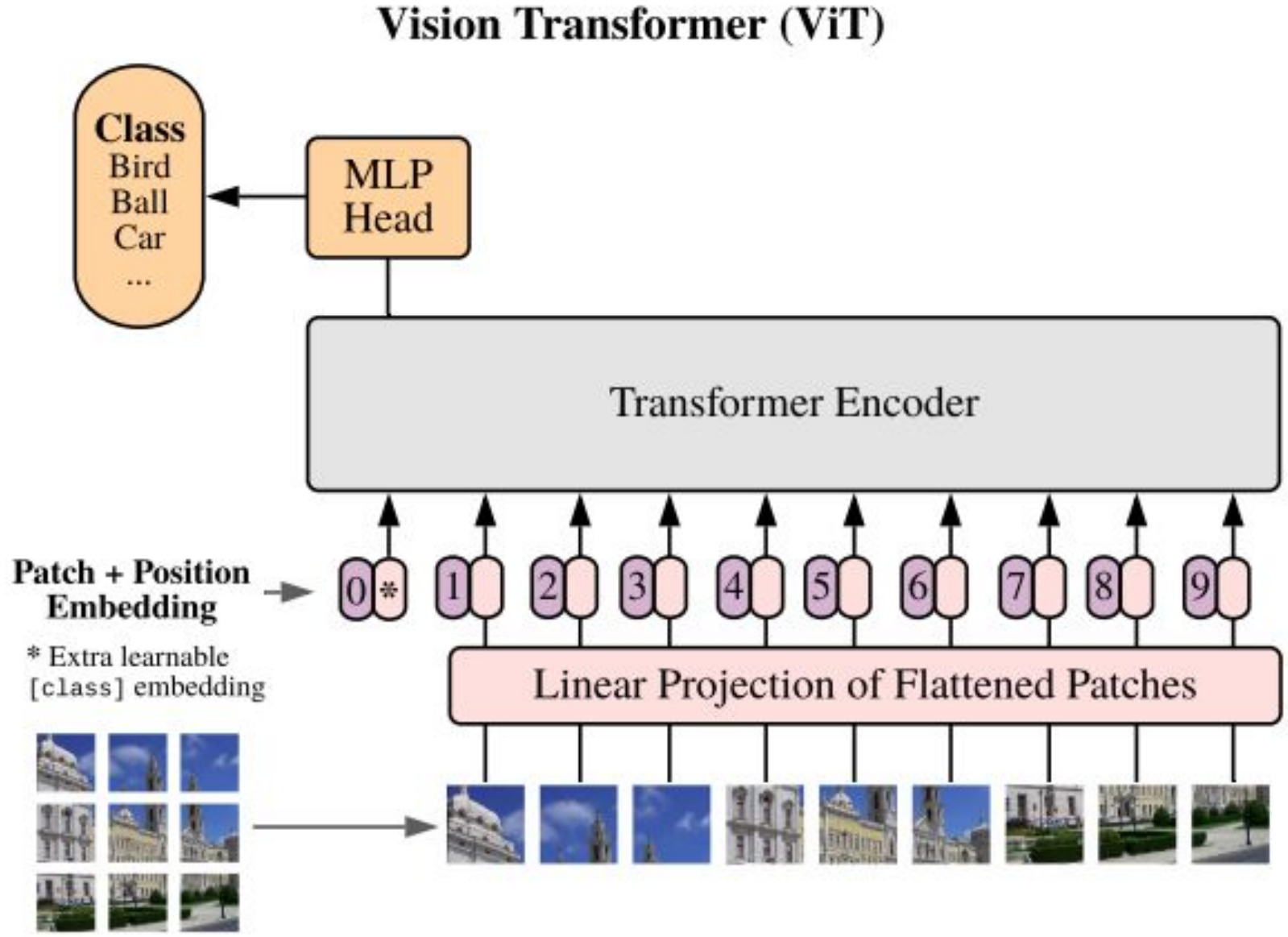


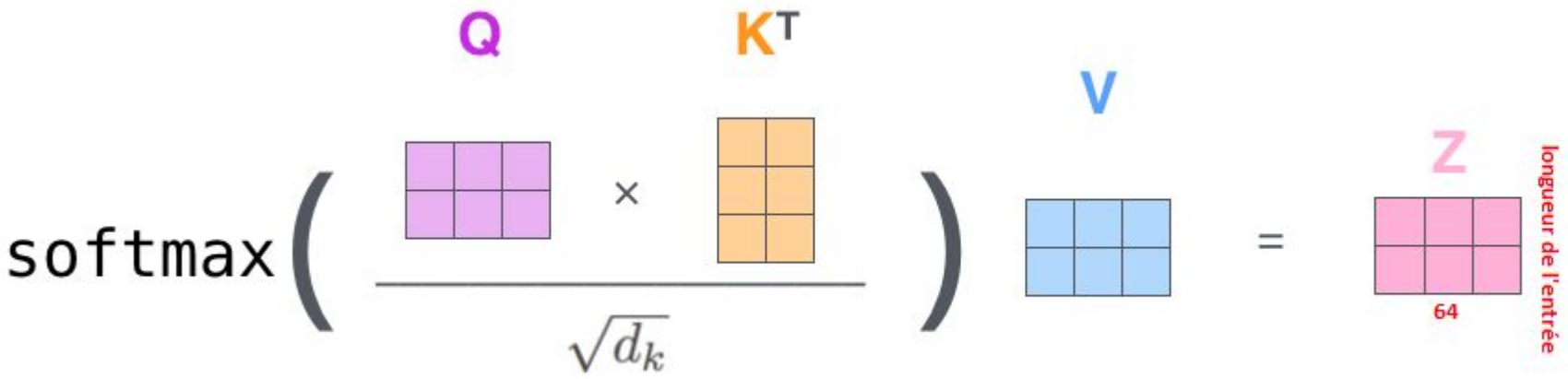


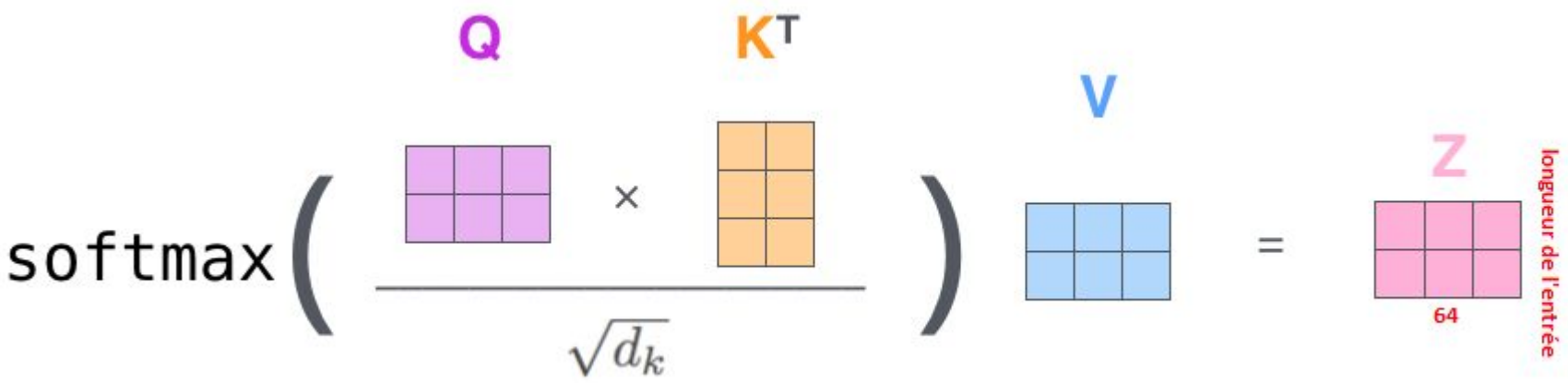
# Architecture transformer



# Architecture transformer



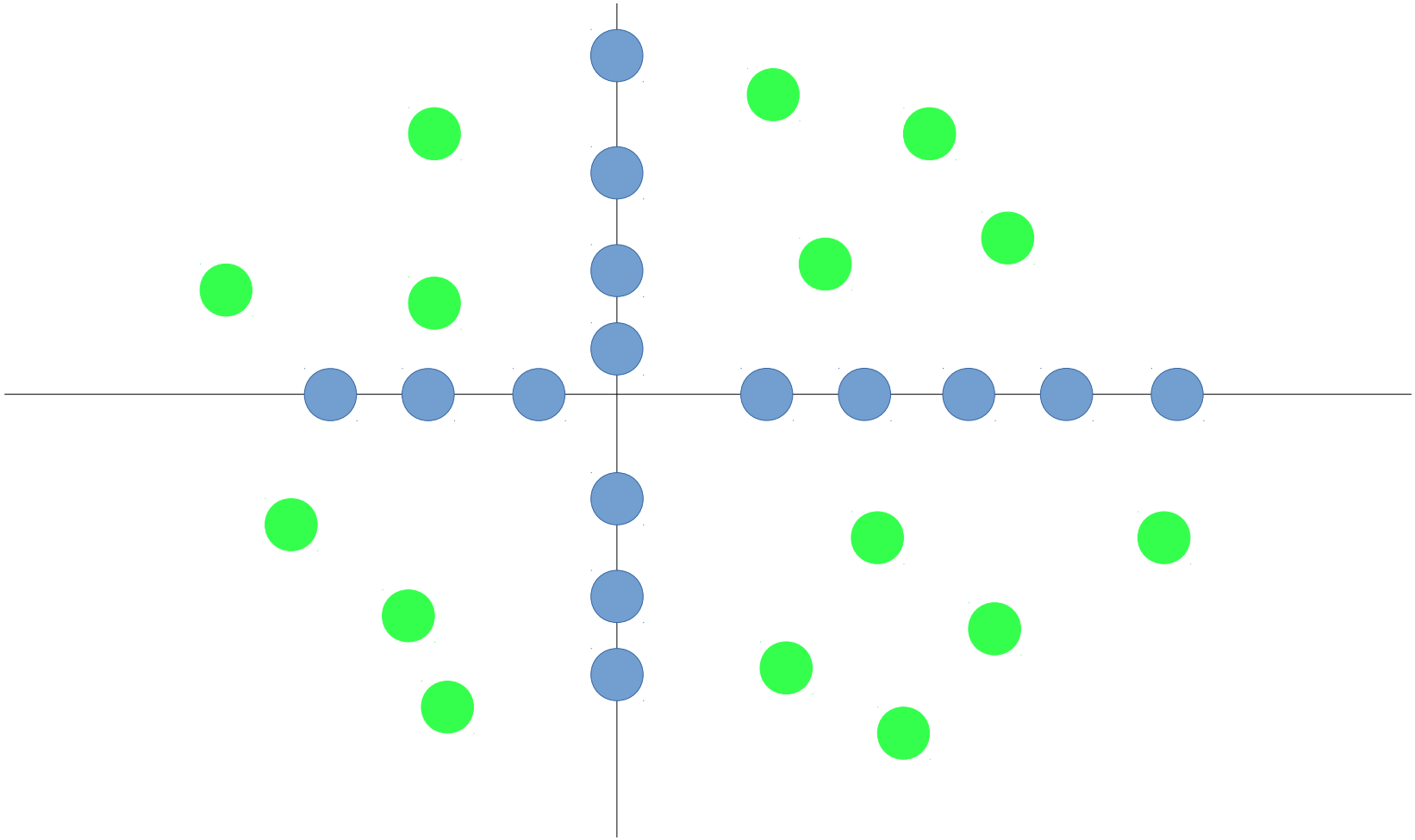
# Architecture transformer



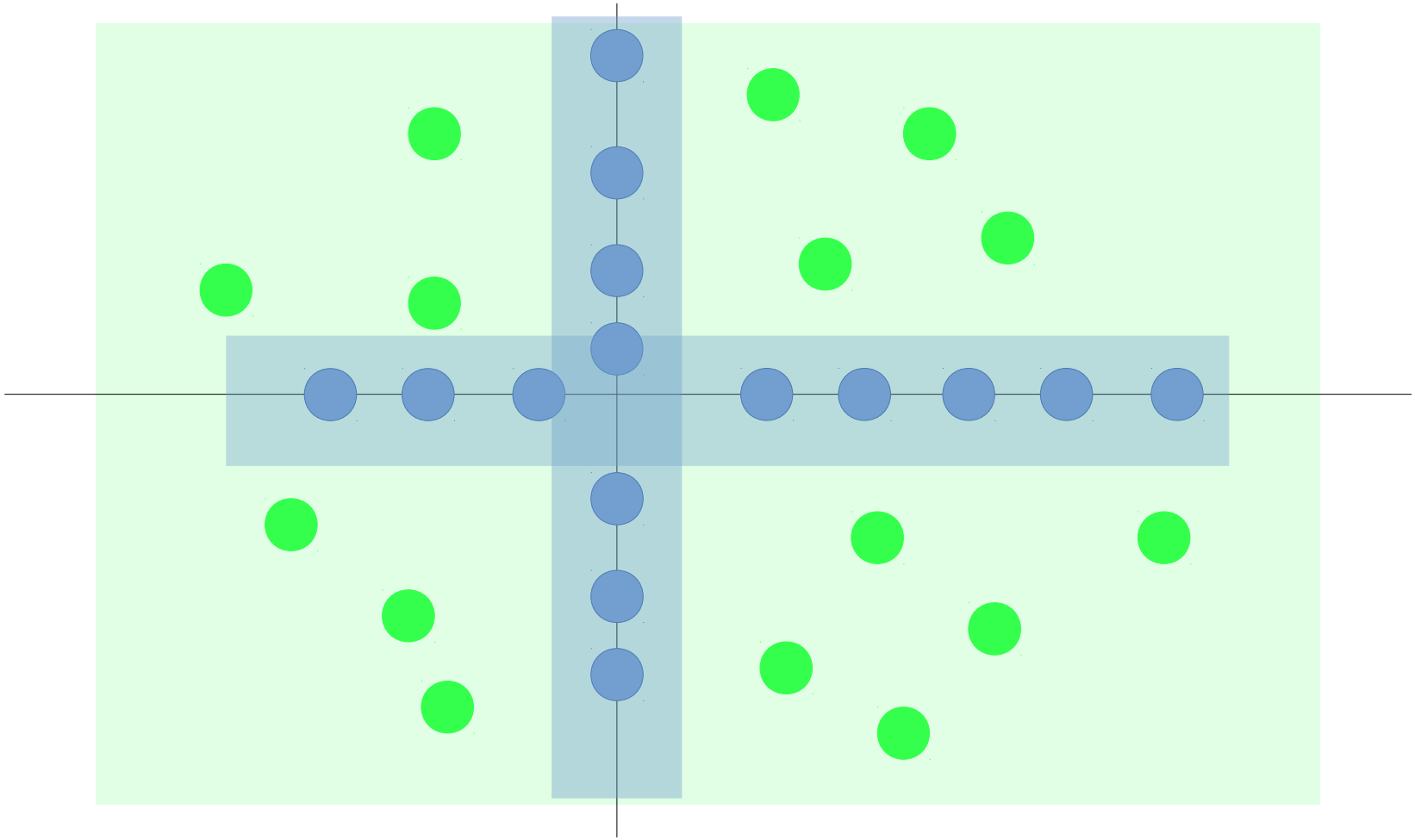

→ Est ce qu'il est vraiment pertinent de décomposer une image en mot visuel ?

**Est ce que les tranformers ont réussi à introduire une multiplication dans une machine additive ?**

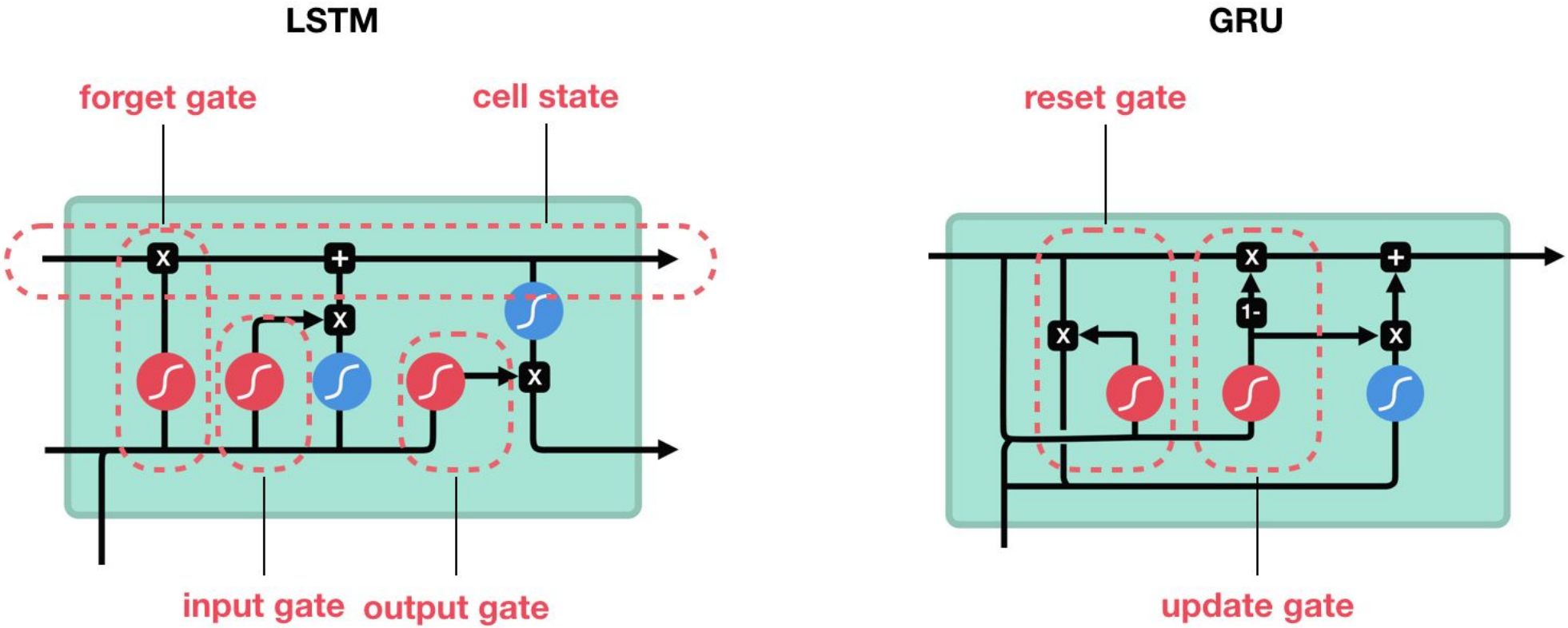
# Architecture transformer



# Architecture transformer



# Architecture transformer



# Limites et perspectives

→ certification / IA de confiance

→ architecture transformer

# Pré-apprentissage

Frugal learning

Self supervised learning

Strong IA

# Pré-apprentissage

*Les algorithmes d'apprentissage nécessitent des milliers d'exemples*

*Alors qu'un humain est capable de comprendre un concept rien qu'avec une description textuelle*

# Pré-apprentissage

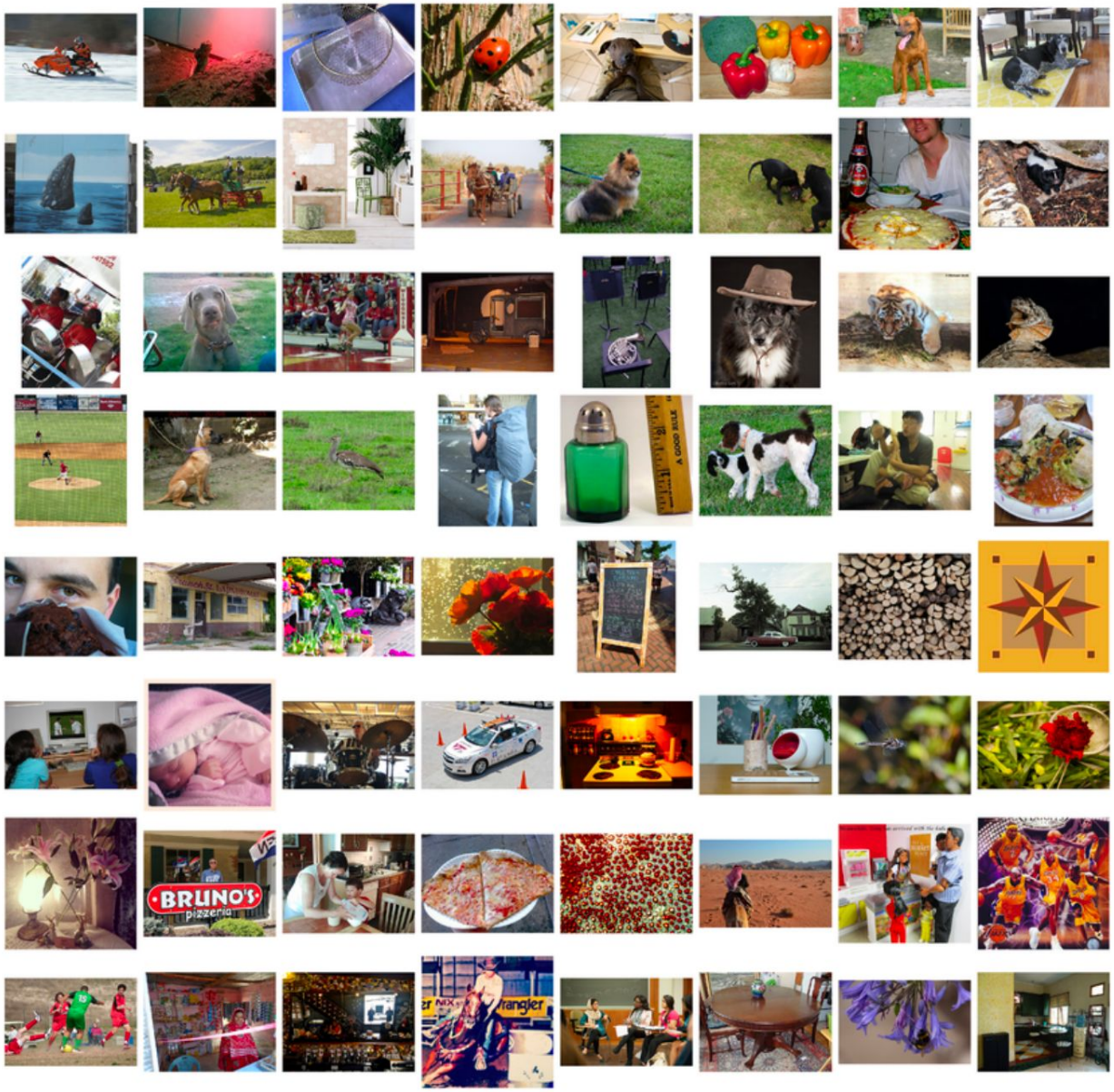
*Les algorithmes d'apprentissage nécessitent des milliers d'exemples*

*Alors qu'un humain est capable de comprendre un concept rien qu'avec une description textuelle*

C'est oublié que les humains ont eu une enfance !

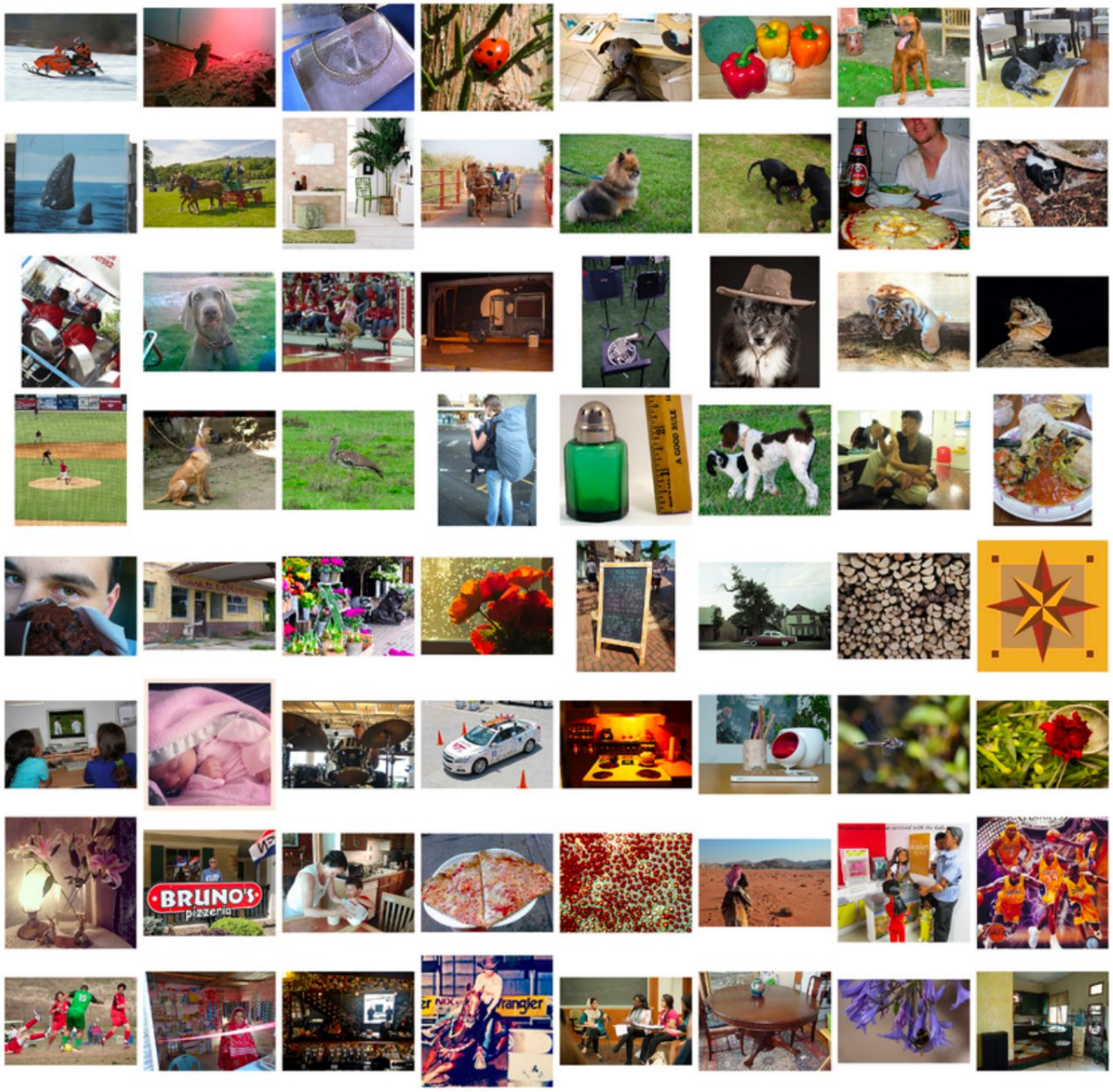
→ **Est ce que bien apprendre n'est pas juste avoir bien pré-appris ??**

# Pré-apprentissage




Imagenet

# Pré-apprentissage




Imagenet

Remote sensing

# Pré-apprentissage

Au delà des performances,
la grande nouveauté des réseaux de neurones est
la possibilité d’utiliser des bouts de réseaux pré-appris sur une nouvelle tache

# Pré-apprentissage

Au delà des performances,
la grande nouveauté des réseaux de neurones est
la possibilité d'utiliser des bouts de réseaux pré-appris sur une nouvelle tache

pré-apprentissage Imagenet → réduction par 10 de la durée d'apprentissage

# Pré-apprentissage

Au delà des performances,
la grande nouveauté des réseaux de neurones est
la possibilité d'utiliser des bouts de réseaux pré-appris sur une nouvelle tache

pré-apprentissage Imagenet → réduction par 10 de la durée d'apprentissage

Les poids ne sont pas adaptés à la nouvelle tache mais « communiquent ».

# Pré-apprentissage

Au delà des performances,
la grande nouveauté des réseaux de neurones est
la possibilité d'utiliser des bouts de réseaux pré-appris sur une nouvelle tache

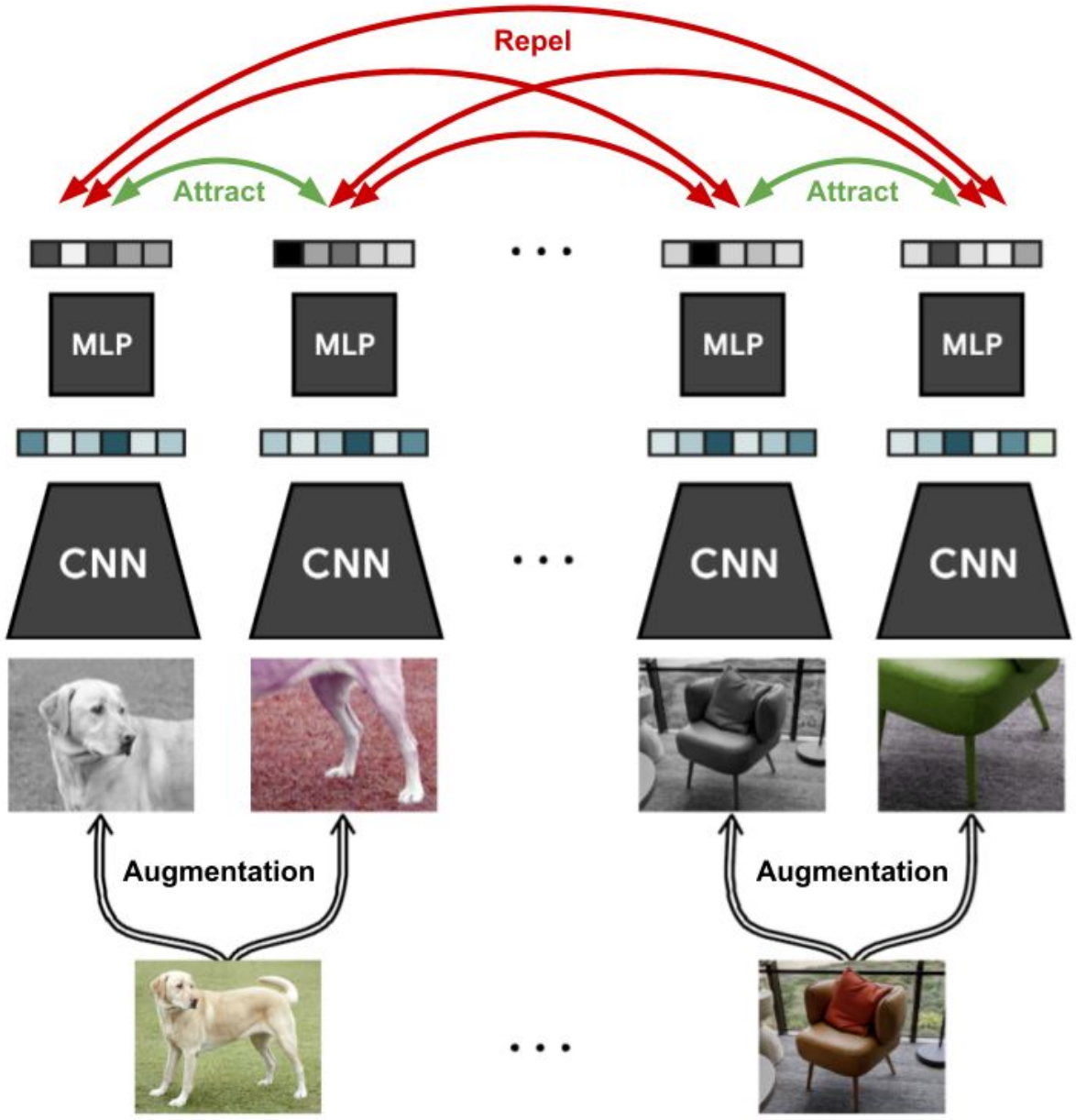
pré-apprentissage Imagenet → réduction par 10 de la durée d'apprentissage

Les poids ne sont pas adaptés à la nouvelle tache mais « communiquent ».

**Green IA, Frugal learning, Few shot learning, 0 shot learning**

→ **trouver le bon préapprentissage**

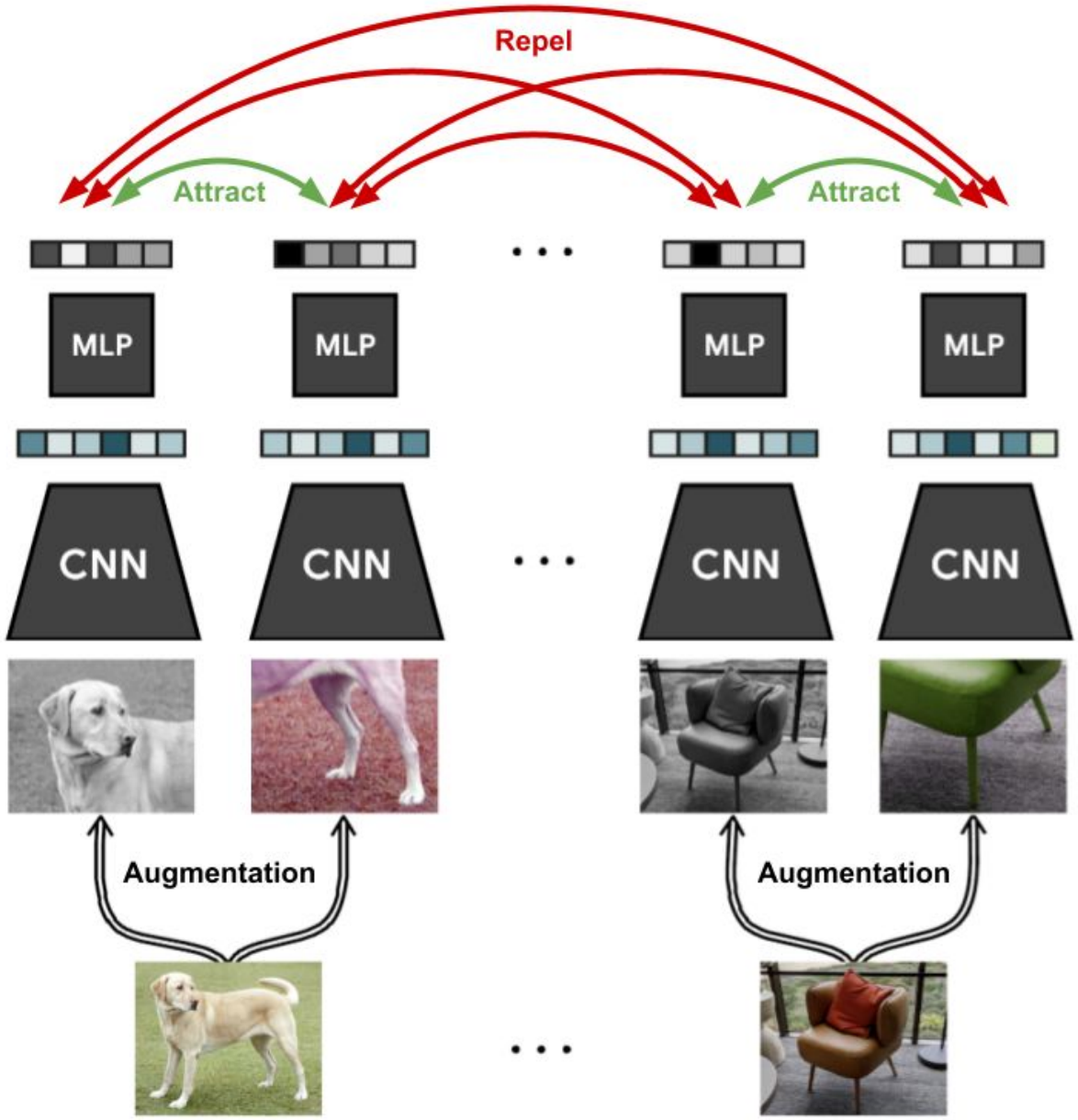
# Pré-apprentissage




Self supervised learning

→ Contrastive learning

# Pré-apprentissage

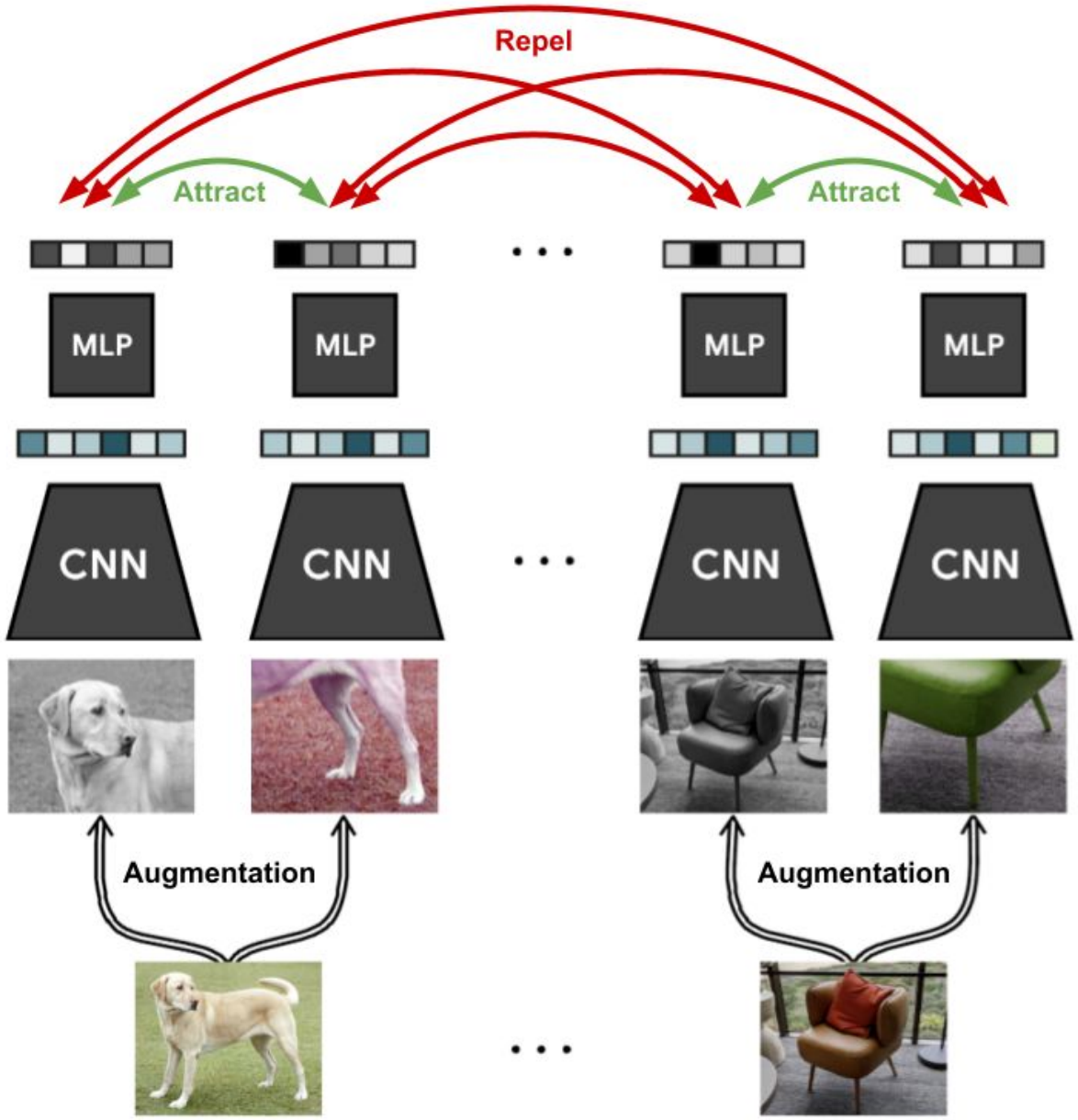



Self supervised learning

→ Contrastive learning

→ performance d'Imagenet avec seulement 10 % des labels

# Pré-apprentissage



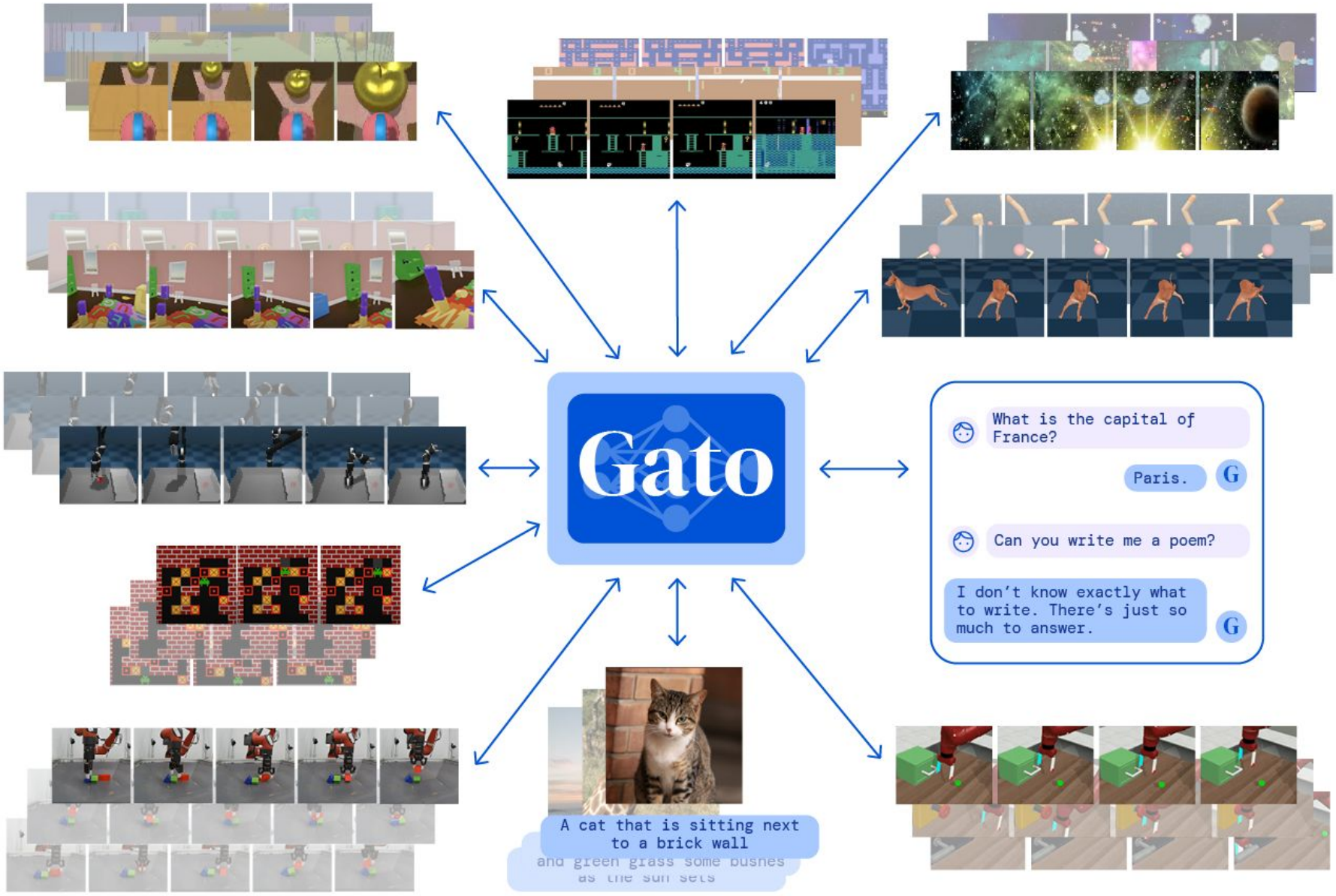

Self supervised learning

→ Contrastive learning

→ performance d'Imagenet avec seulement 10 % des labels

→ augmentation des performances supervisées !

# Pré-apprentissage



# Pré-apprentissage

Deux chemins vers
l’IA forte

Préapprentissage fort

Préapprentissage faible
Symbolique et numérique
Hybridation
IA classique et apprentissage

# Limites et perspectives

→ certification / IA de confiance

→ architecture transformer

→ préapprentissage et chemin vers l’IA forte