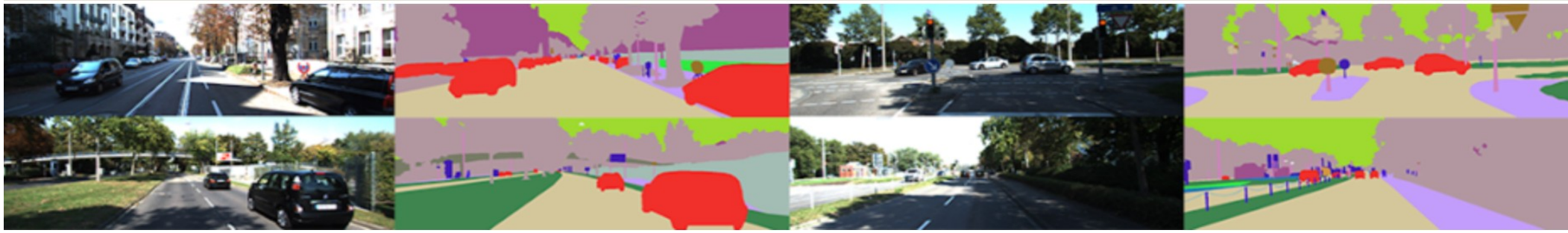
Segment Anything - SAM

Méthodes Avancées en Segmentation d'Images

Robert Maria

Strasbourg
23.02.2024

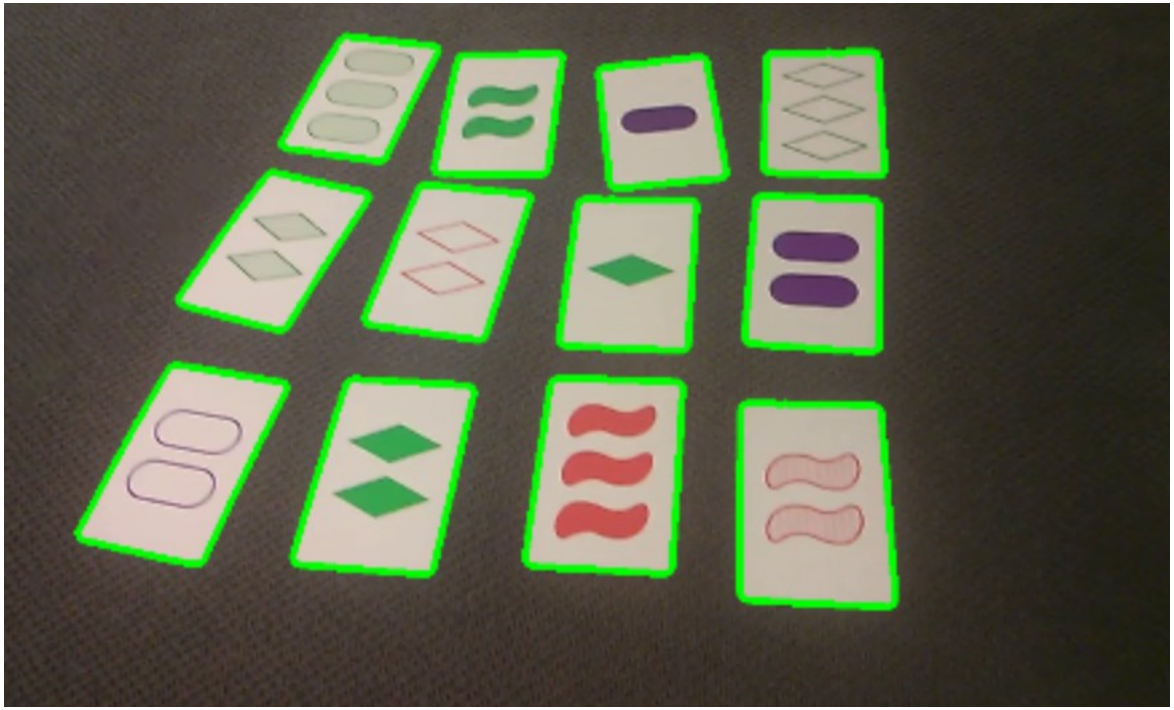
La segmentation d'image



- diviser une image en parties ou segments

La segmentation d'image

1. Détection de Contours

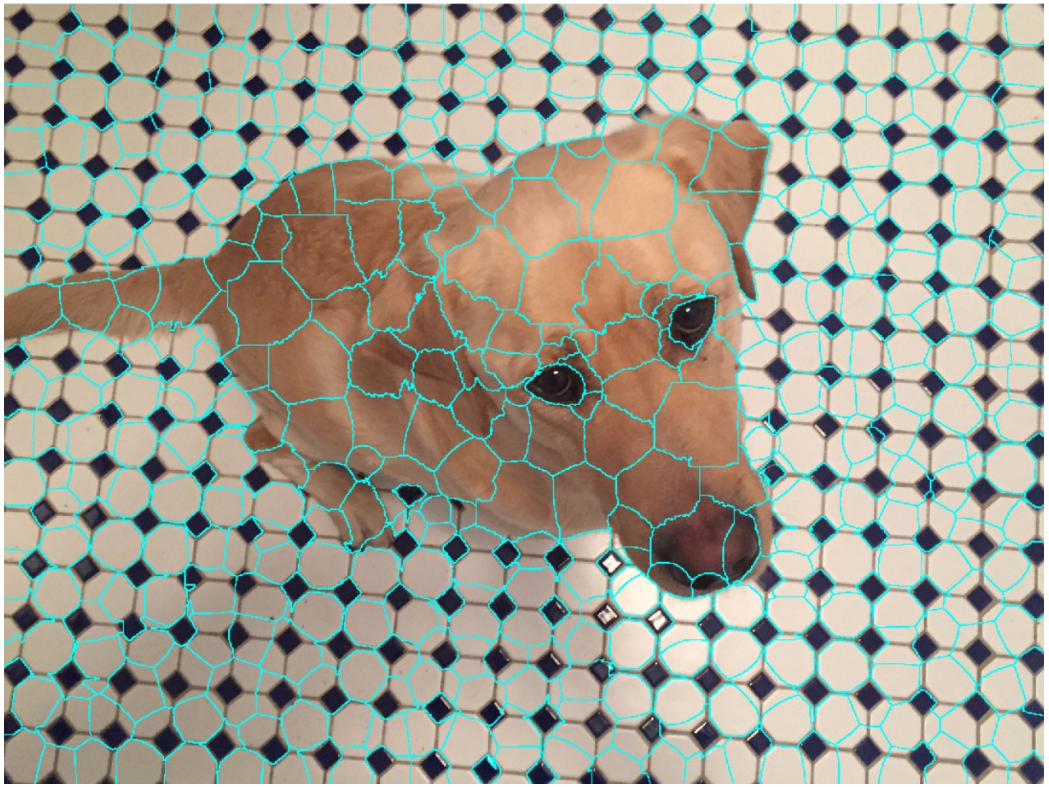


- étape préliminaire dans de nombreuses tâches de segmentation

- délimiter les frontières avant un traitement plus approfondi

La segmentation d'image

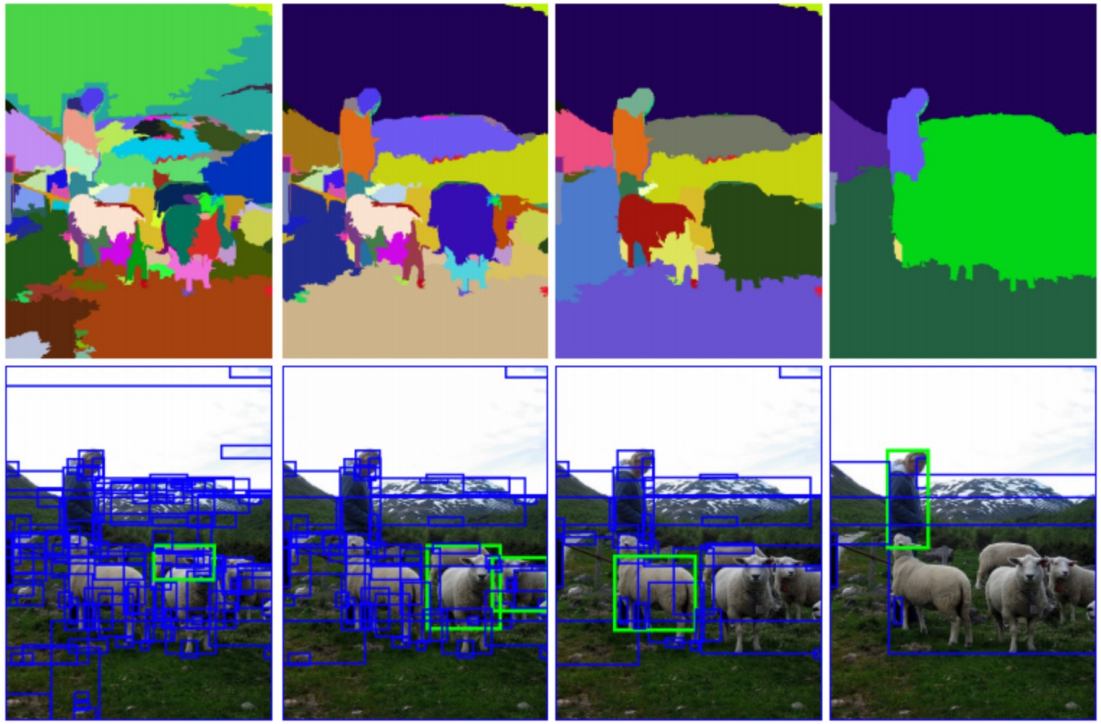
2. La superpixelisation



- regrouper des pixels voisins (ayant des caractéristiques similaires) en unités plus grandes appelées superpixels.

La segmentation d'image

3. La proposition d'objets ("object proposal")

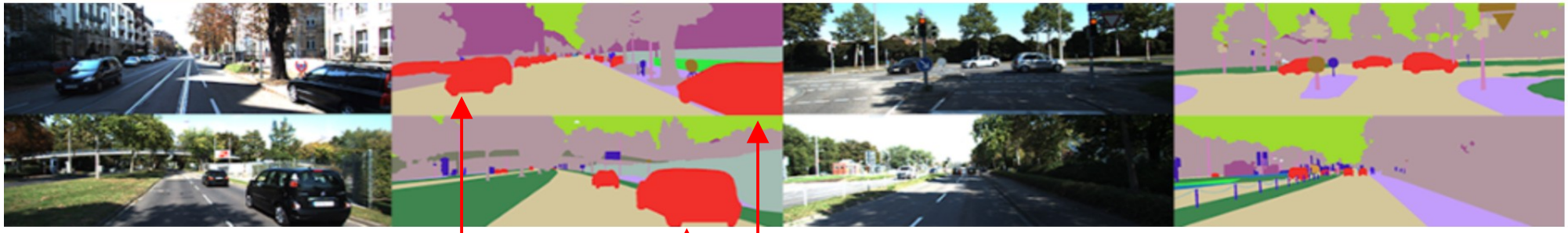


- génération de régions candidates

- Ex: selective search

La segmentation d'image

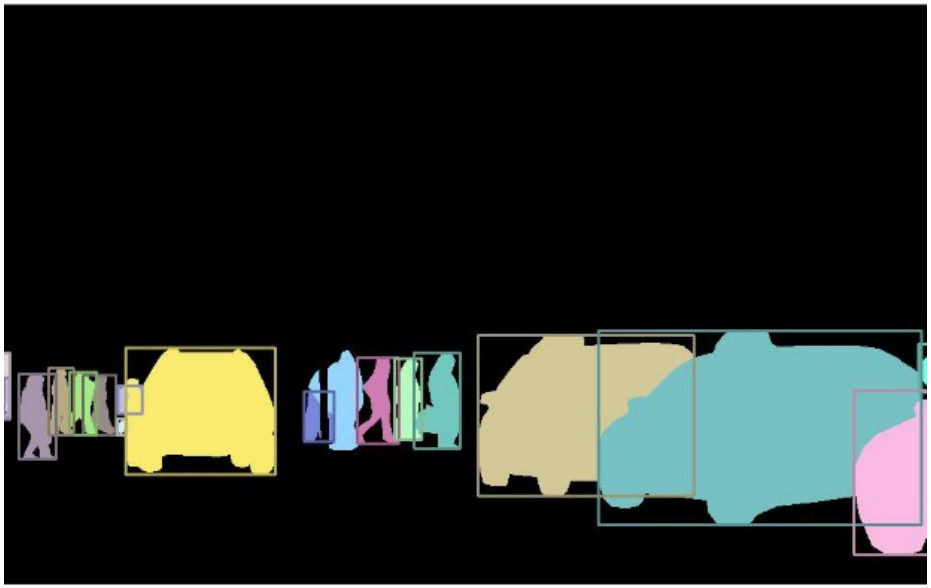
4. Segmentation Sémantique



chaque pixel de l'image est classé dans une catégorie spécifique

La segmentation d'image

5. Segmentation par Instance



- distinguer entre différentes instances de la même classe

La segmentation d'image

5. Segmentation Panoptique

- Segmentation par instance + Segmentation sémantique

Objectifs - SAM

Zero-shot learning

Classes non vues lors de l'entraînement

Few-shots learning

Petit nombre d'exemples (1 ou 2)
pour chaque classe

Flexible prompting

Points, rectangles, texte

Flexibilité à l'ambiguïté

Un nouveau modèle de fondation

source

Modèles de fondation

Succès = f(tâche, modèle, données)

Modèles de fondation

Succès = f(tâche, modèle, données)

tâche = f(**energie**, **calcul**, **mémoire**, **temps**)

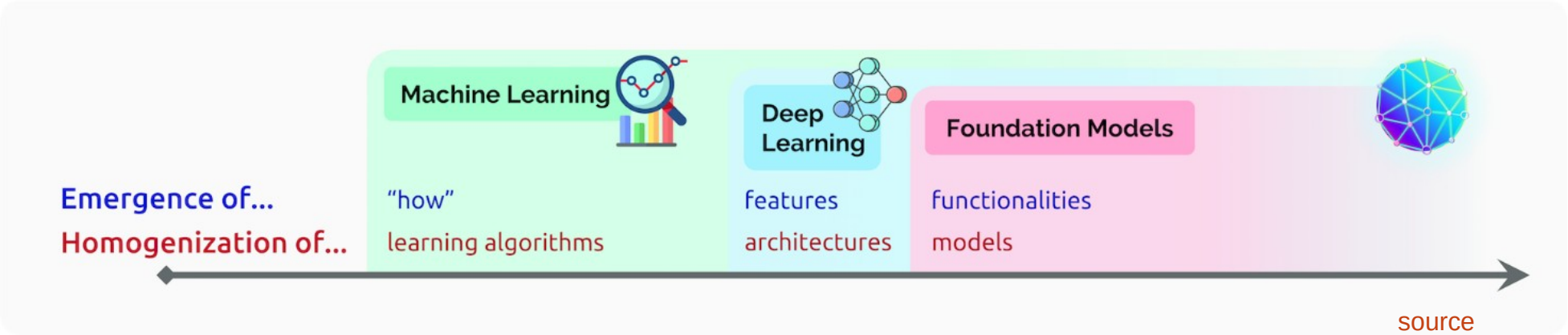
tâche = meilleure précision avec la **charge de ressources**

change avec le temps

Modèles de fondation

L'IA connaît un changement de paradigme

émergence de modèles (BERT, DALL-E, GPT-3, CLIP) qui peuvent être adaptés à une vaste gamme de tâches en aval.



Traitement d'image classique — CNNs — Transformers

beaucoup — **moins** — **encore moins**

biais inductifs

source

Modèles de fondation

L'émergence

le comportement d'un système est induit implicitement plutôt qu'explicitement construit

L'homogénéisation

consolidation des méthodologies pour construire des systèmes à travers un large éventail d'applications

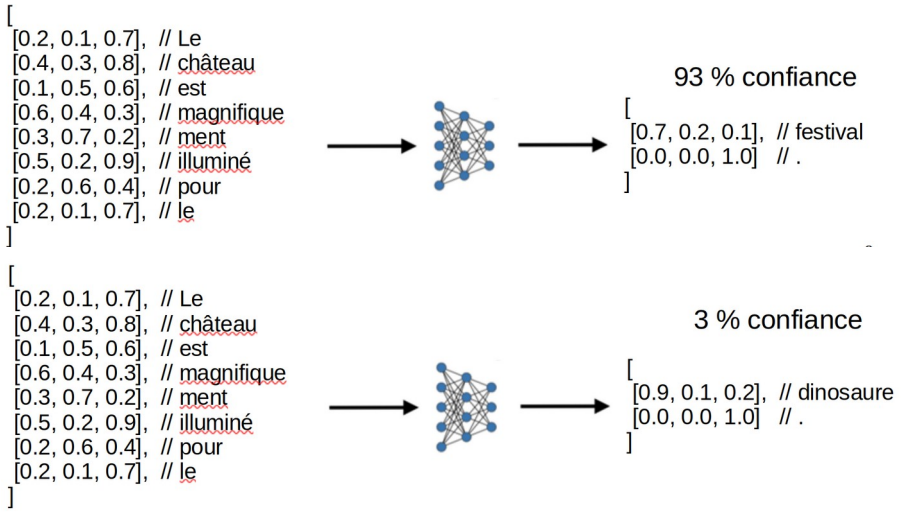
levier important pour de nombreuses tâches

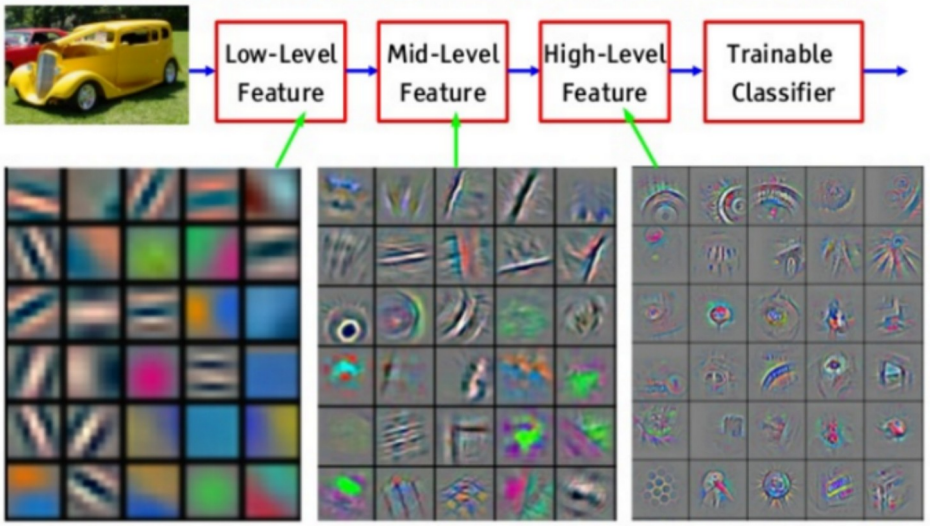
points uniques de défaillance

Modèles de fondation

Succès = f(tâche, modèle, données)

“Modèles de fondation”



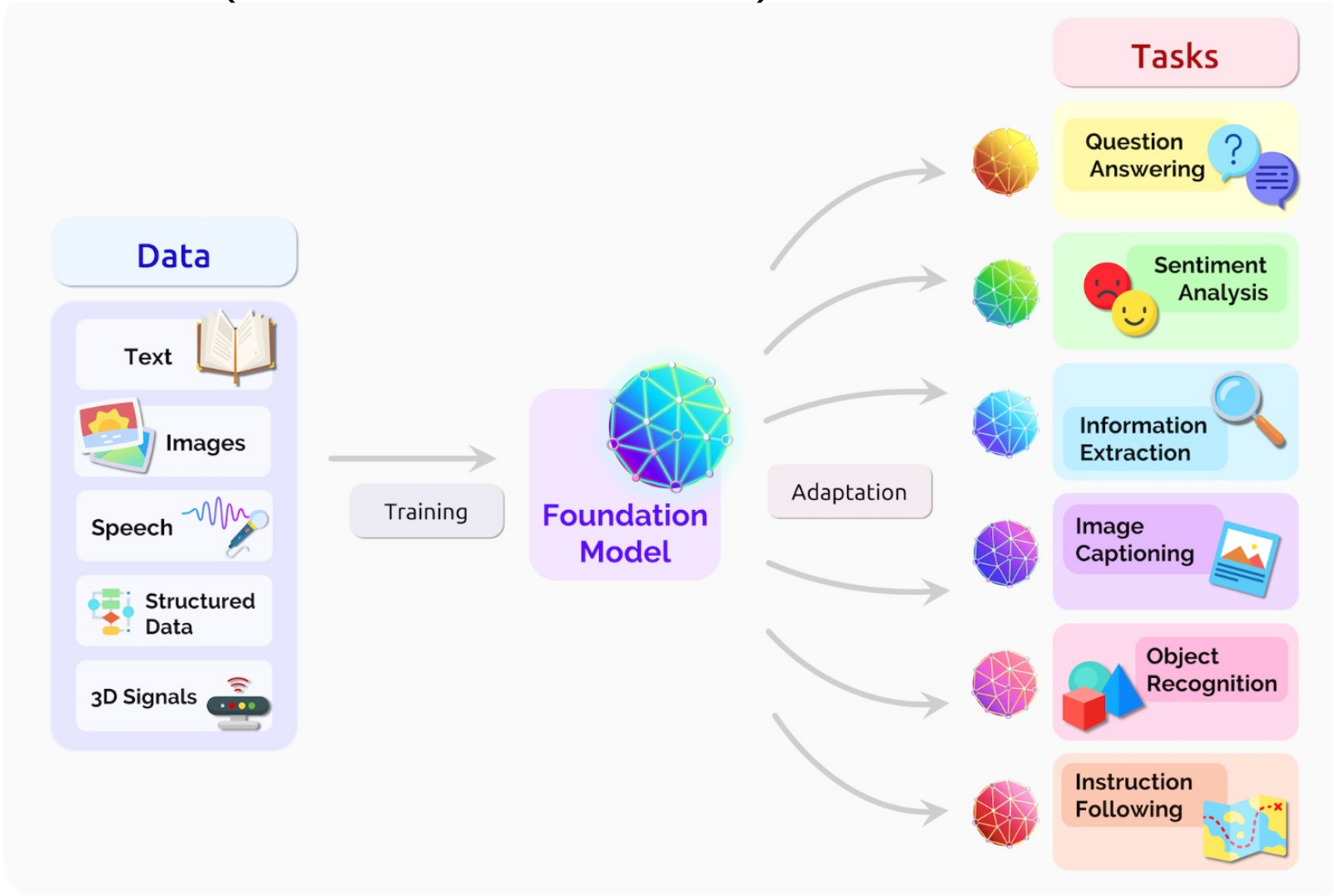


apprentissage par transfert + échelle

préentraînement + fine-tuning

Modèles de fondation

Succès = f(tâche, modèle, données)



source

Segment Anything (SAM)

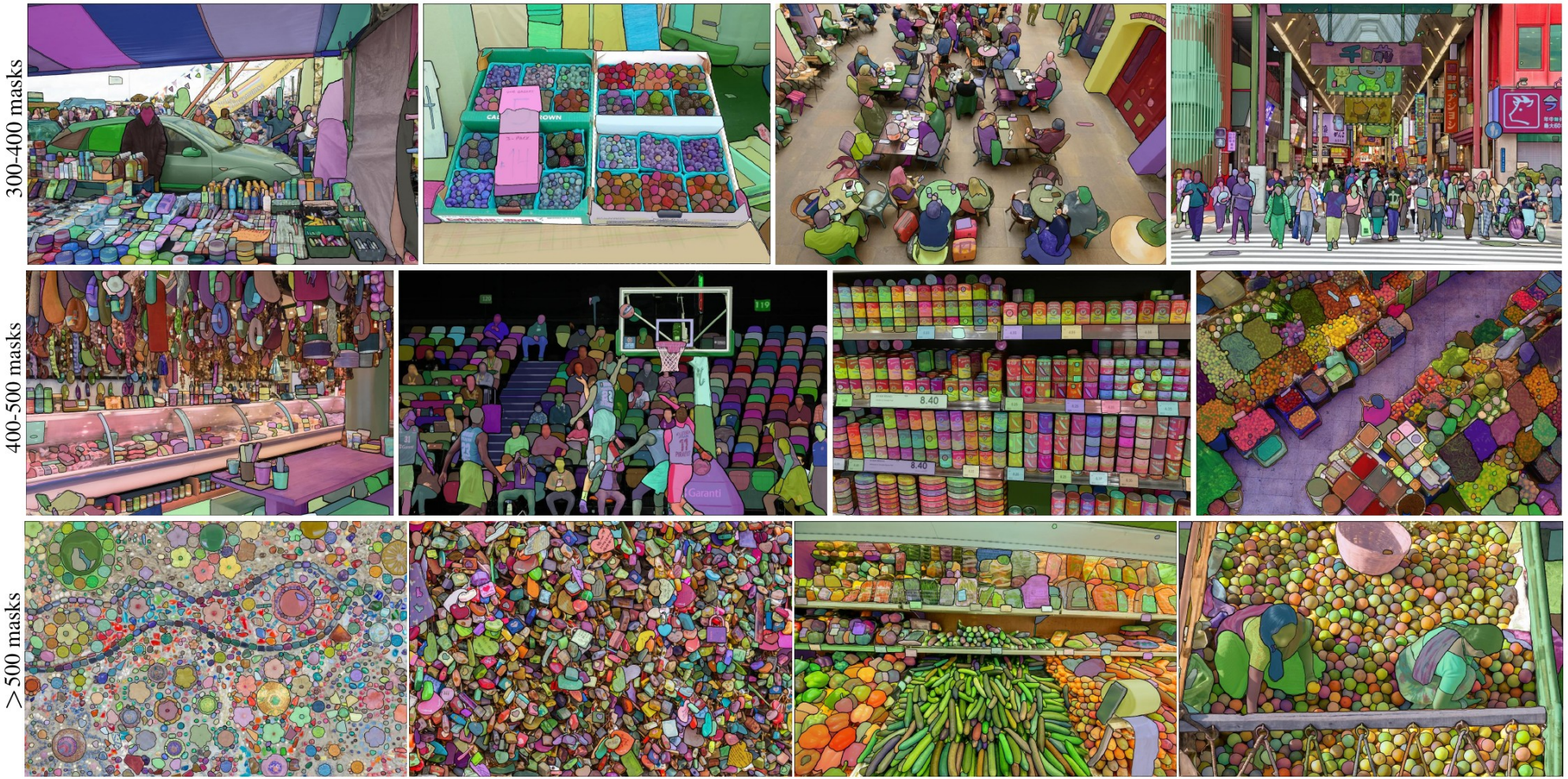
Modèle de fondation dans la segmentation d'images

Segment Anything (SAM)

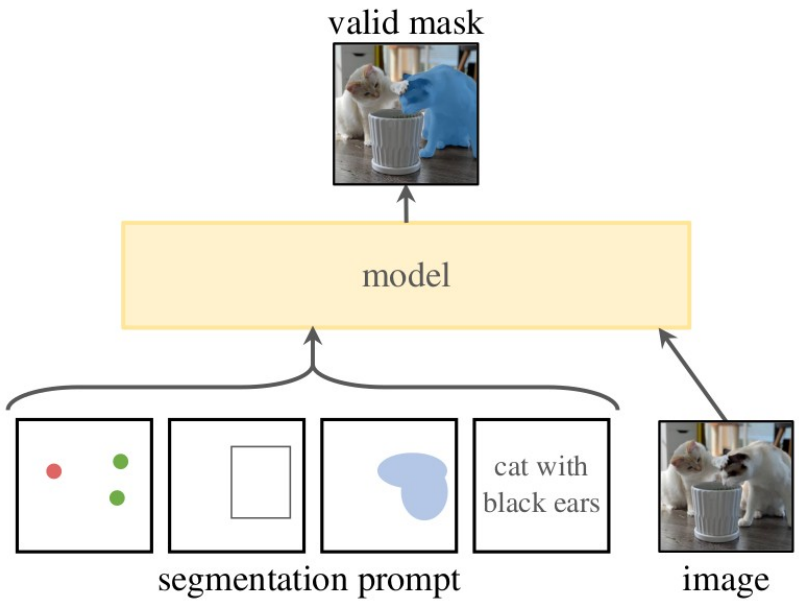
SA-1B dataset release:

Segment Anything (SAM)

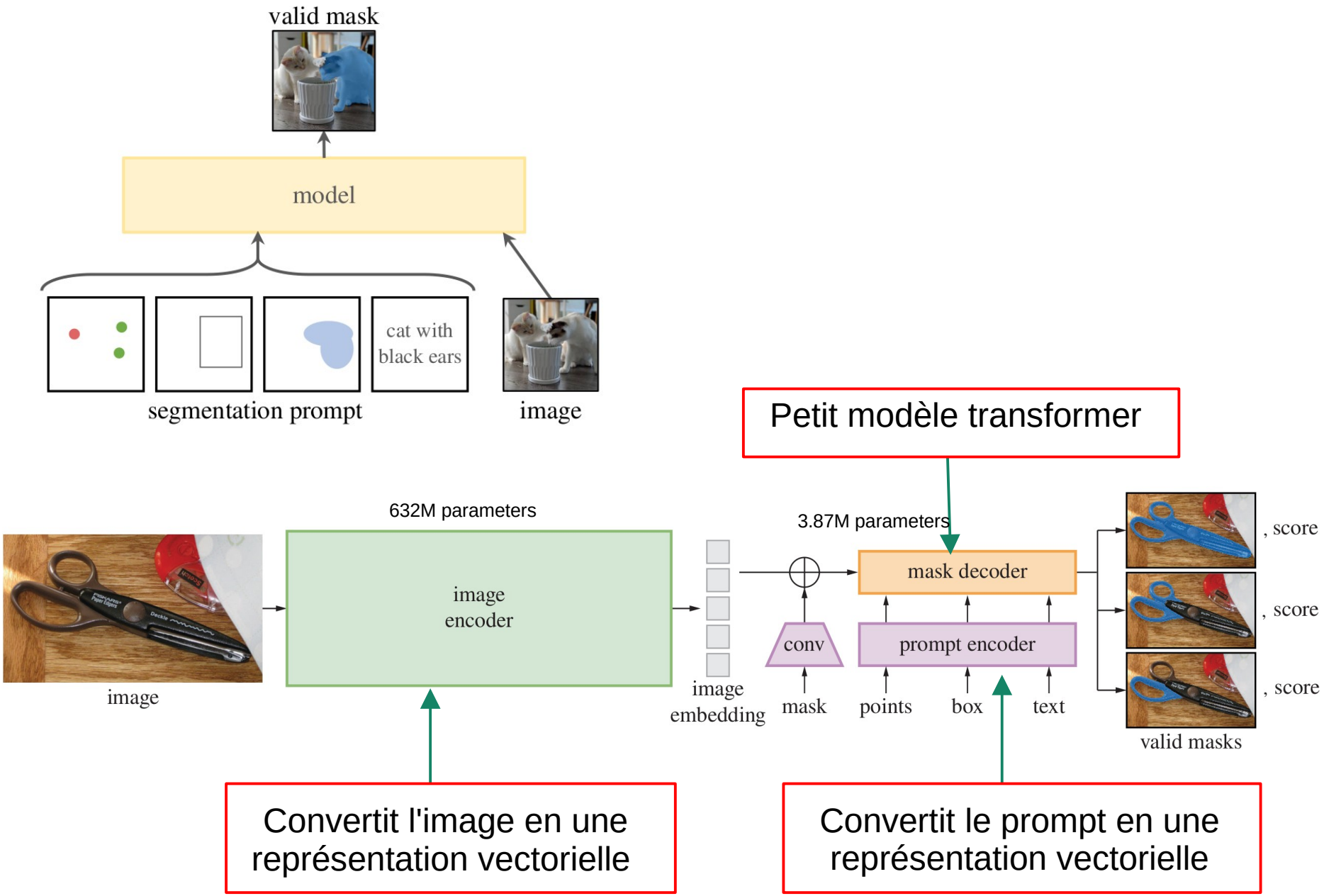
SA-1B dataset release:



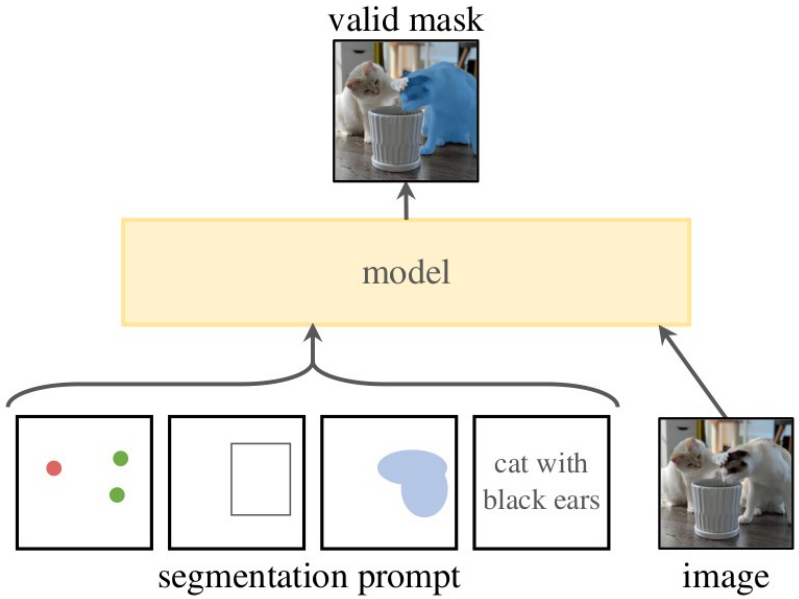
Segment Anything (SAM)



Segment Anything (SAM)



Segment Anything (SAM)

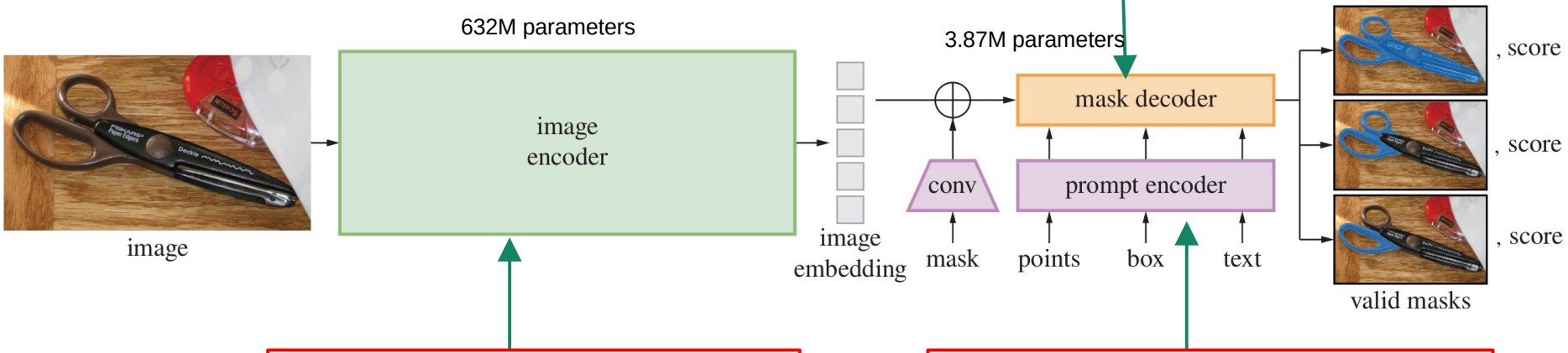


Encodeur d'image = ViT pré-entraîné MAE

ViT = Visual Transformer

MAE = Masked autoencoders

Petit modèle transformer



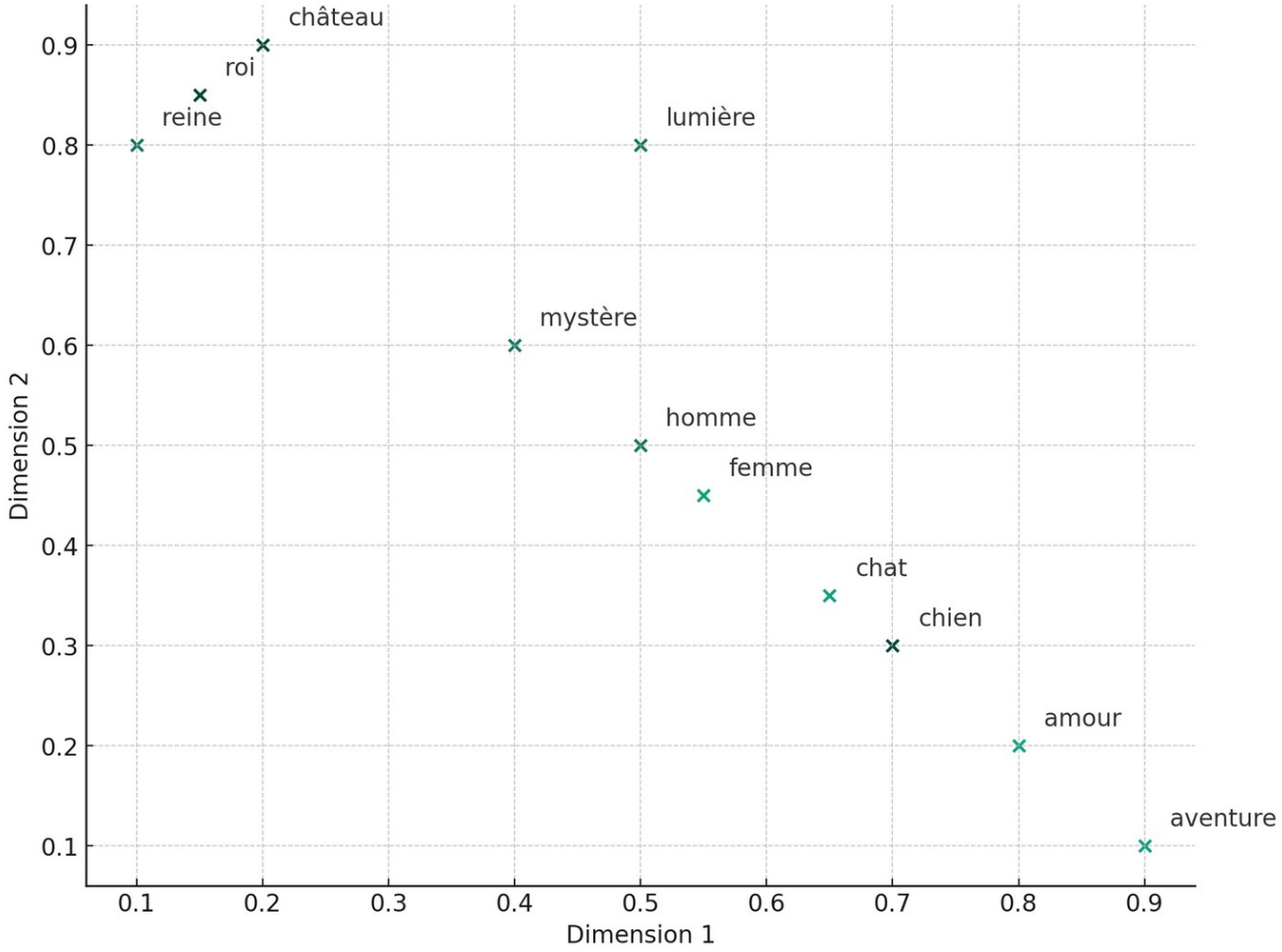
Convertit l'image en une représentation vectorielle

Convertit le prompt en une représentation vectorielle

Embeddings

Embeddings = représentations vectorielles d'images, de texte

Embeddings



Roi – homme + famme = reine

roi = [0.9, 0.1]
homme = [0.5, 0.5]
femme = [0.4, 0.6]
reine = [0.8, 0.2]

roi - homme + femme
= [0.9, 0.1] - [0.5, 0.5] + [0.4, 0.6]
= [0.8, 0.2]

Transformers

Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

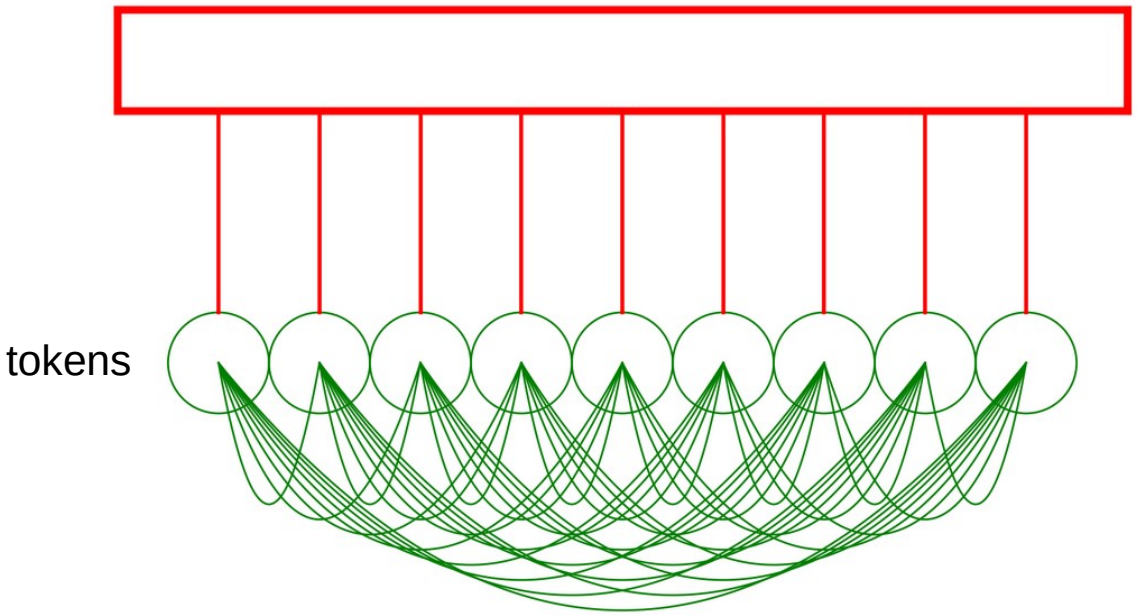
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Tache: **machine translation**

Transformers

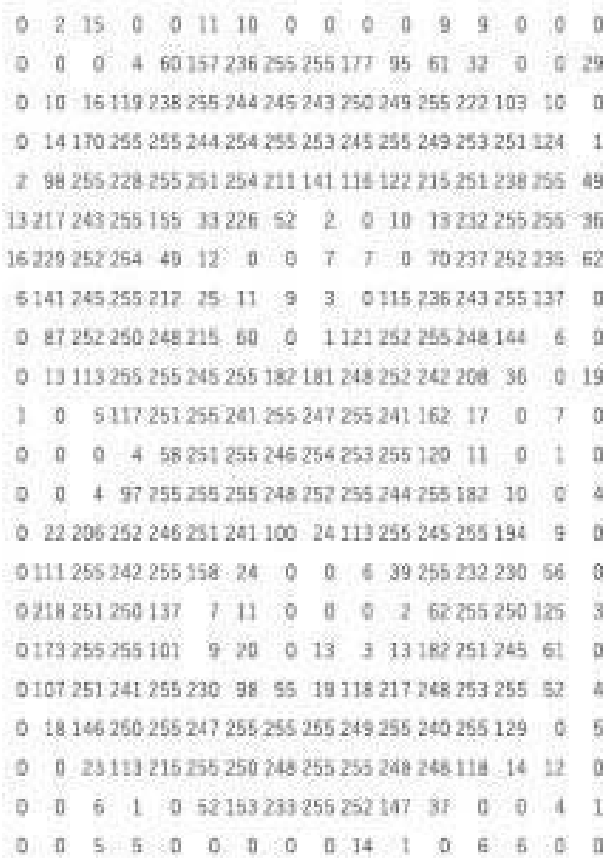
Text



Attention = pairwise inner-product

Quadratic operation

Images



Visual Transformer (ViT)

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

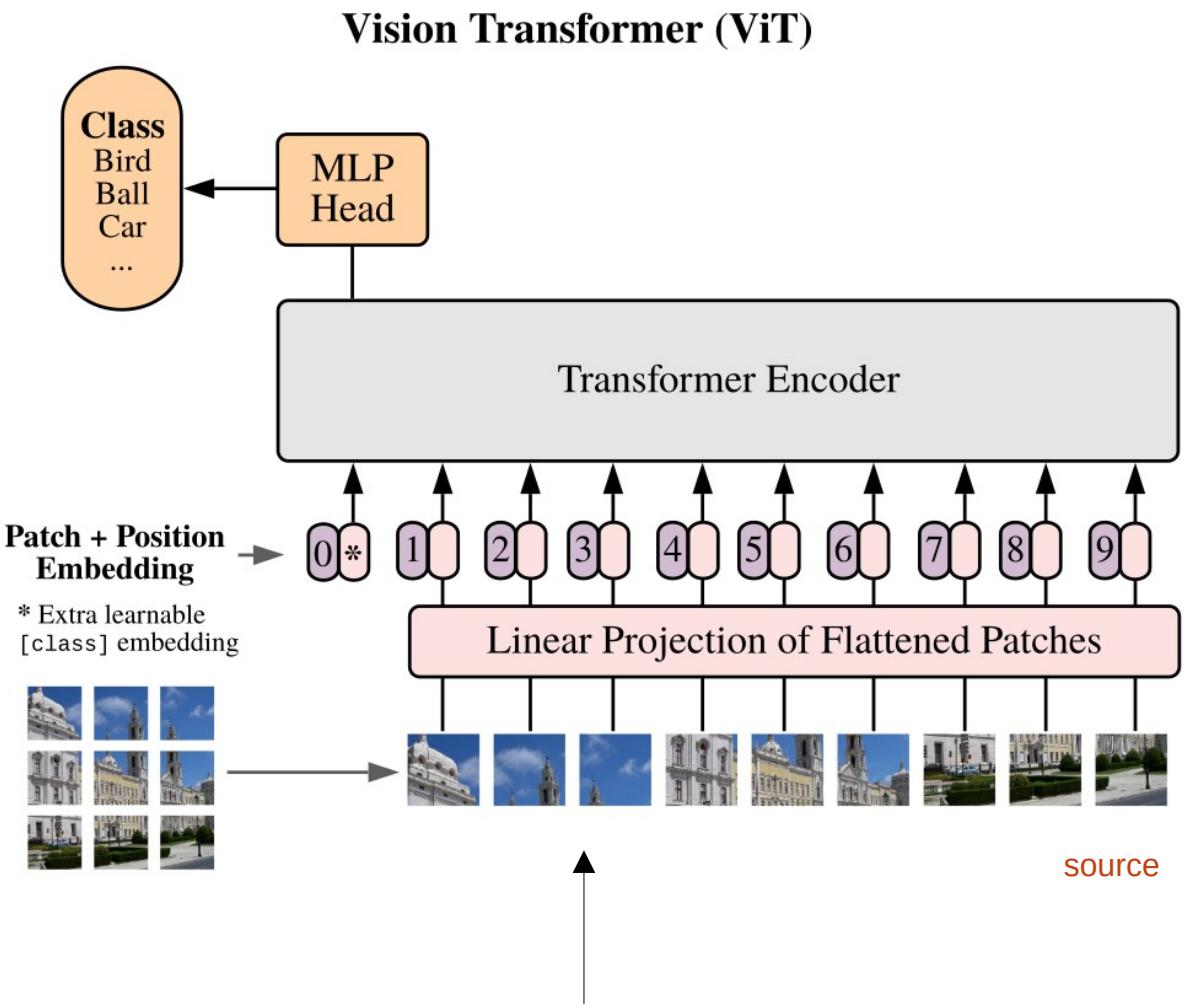
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

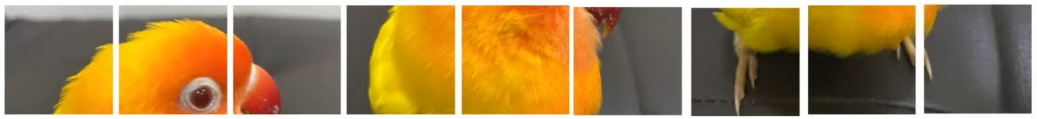
Visual Transformer (ViT)

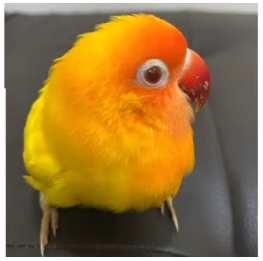


Patchs de 16x16 pixels

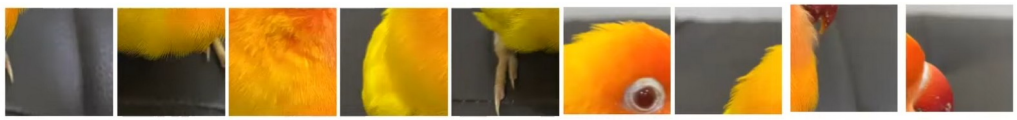
Visual Transformer (ViT)

Raison du codage de position

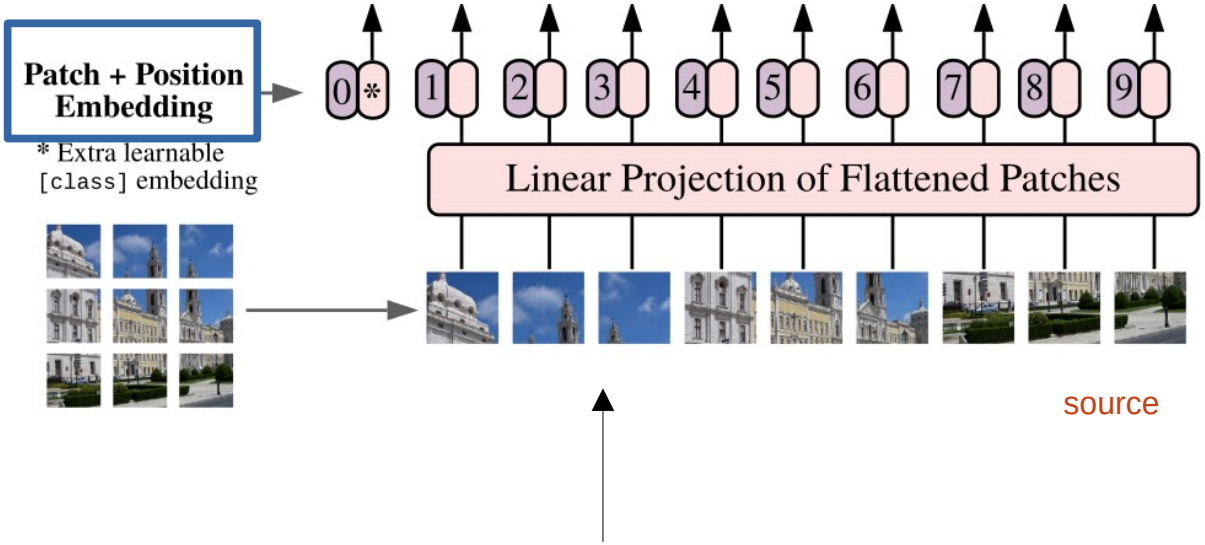




Oiseau



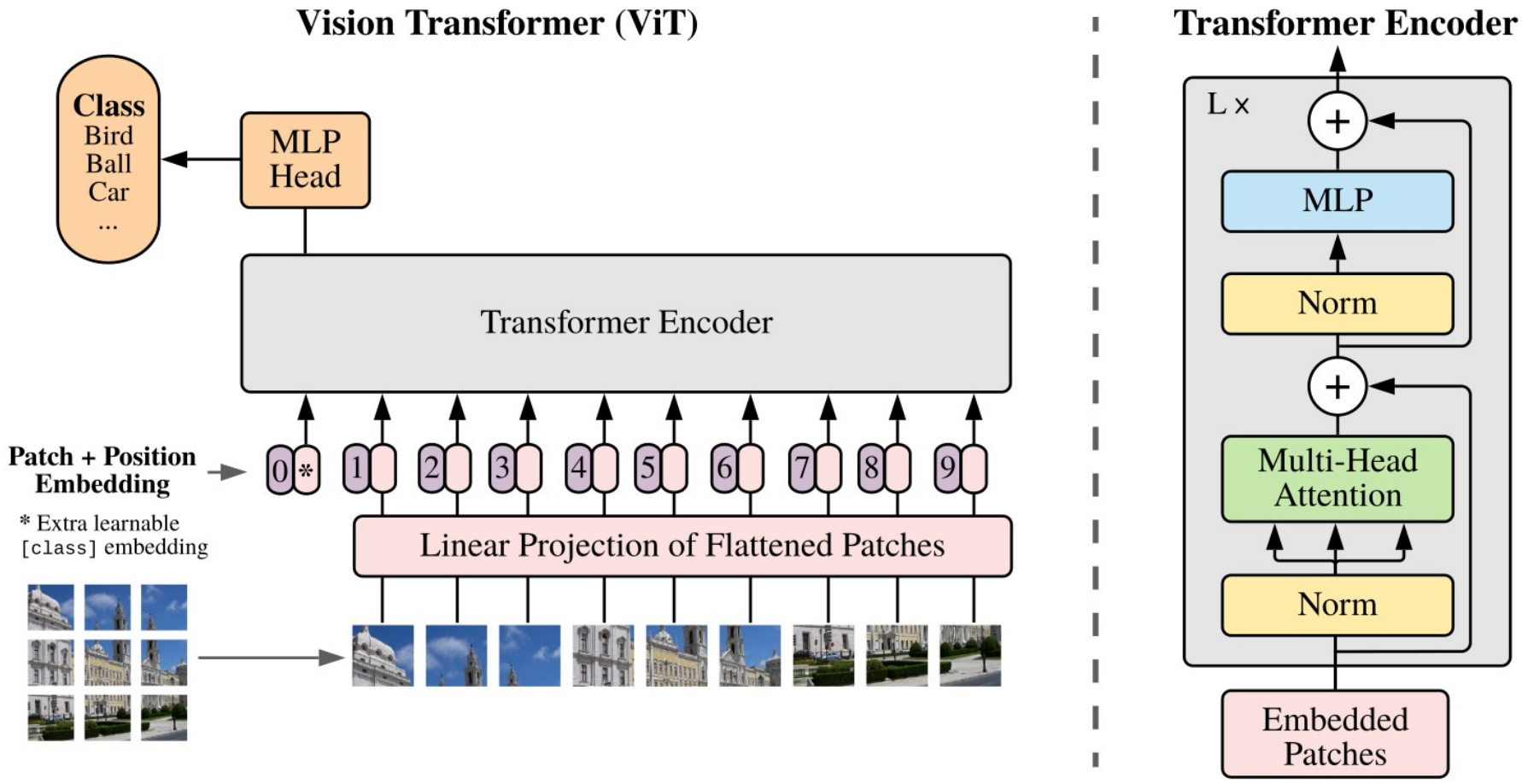
Oiseau ?



source

Patches 16x16 pixels

Visual Transformer (ViT)

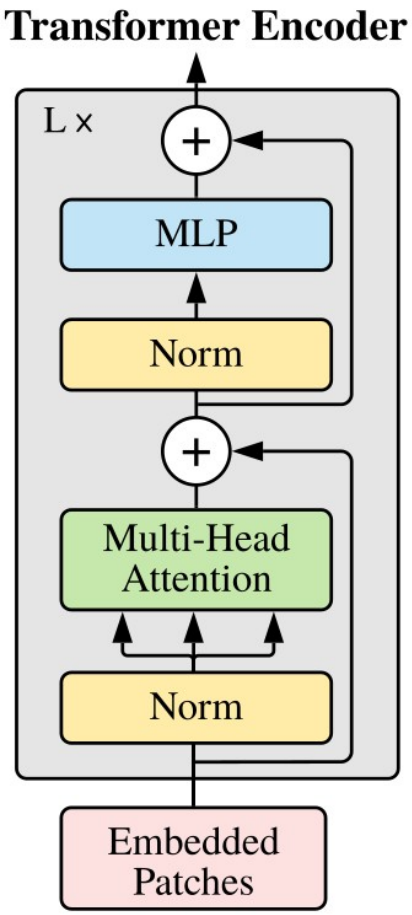


source

Visual Transformer (ViT)

permet aux embeddings de « penser » de manière indépendante

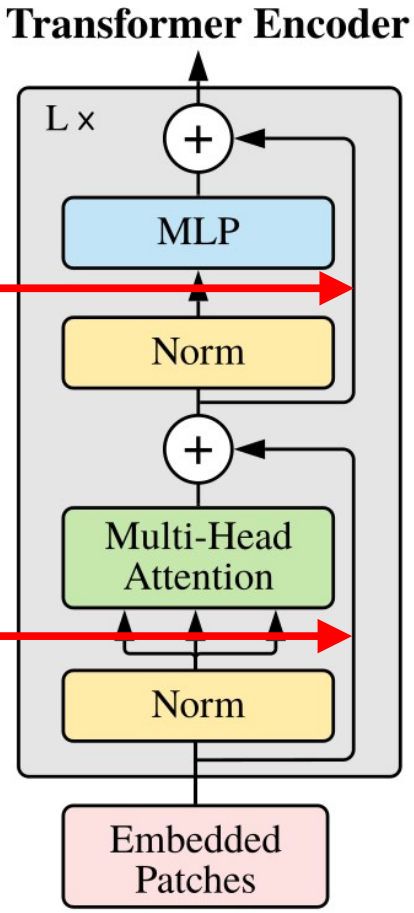
permet aux embeddings de communiquer



source

Visual Transformer (ViT)

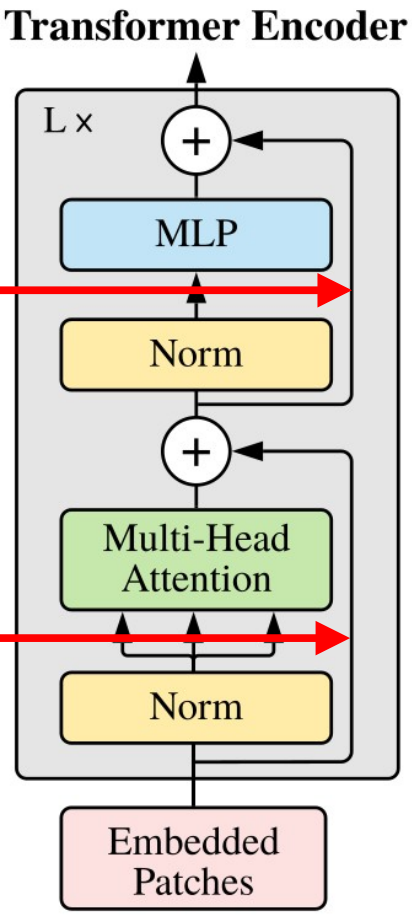
connexions résiduelles



source

Visual Transformer (ViT)

connexions résiduelles
aide à la circulation des gradients

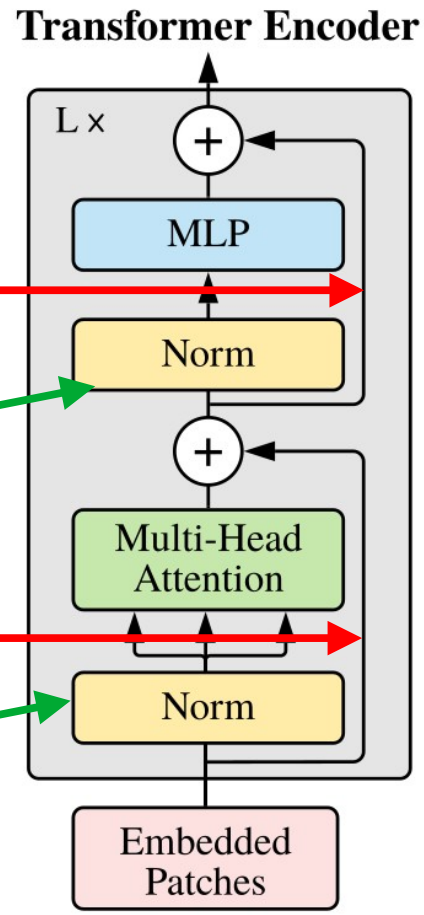


source

Visual Transformer (ViT)

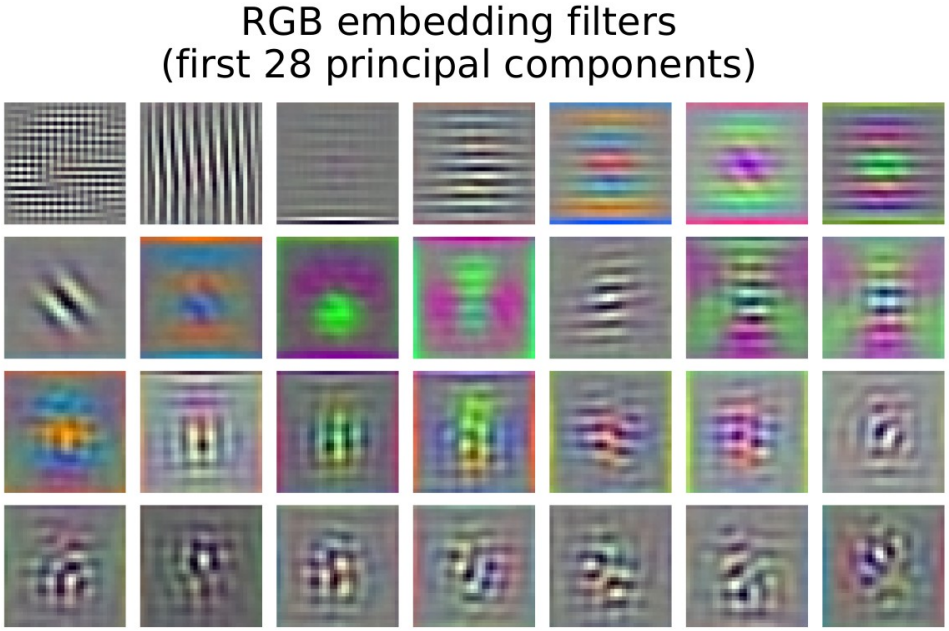
connexions résiduelles
aide à la circulation des gradients

normalisation par couche
stabilise l'apprentissage



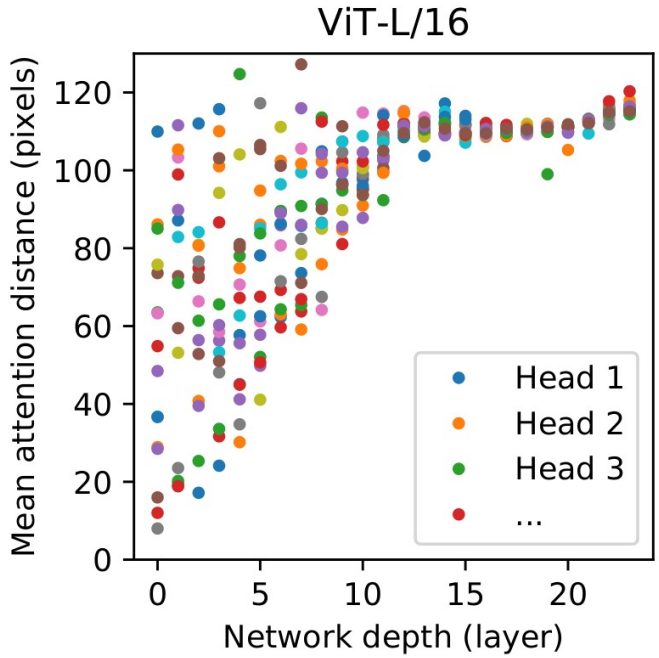
source

Visual Transformer (ViT)



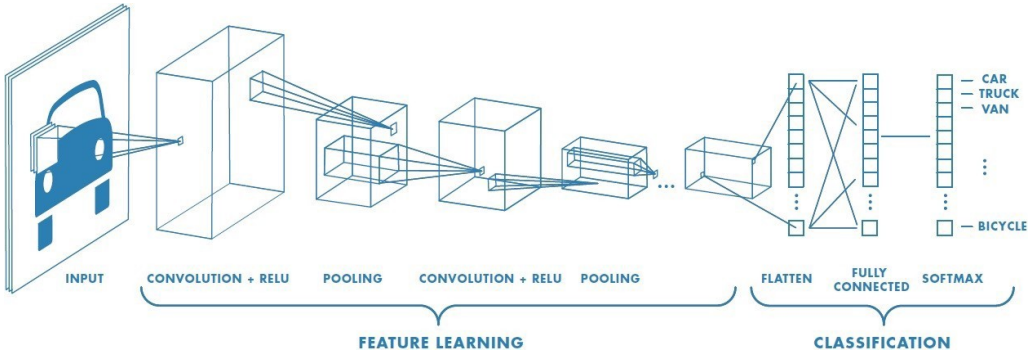
source

Feature maps ressemblant à des CNNs



source

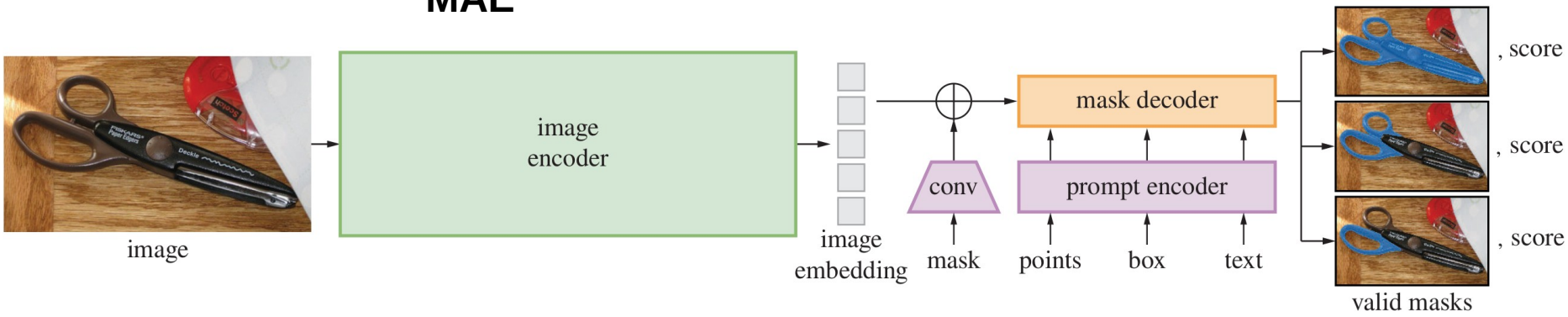
Champs récepteurs globaux dès le début



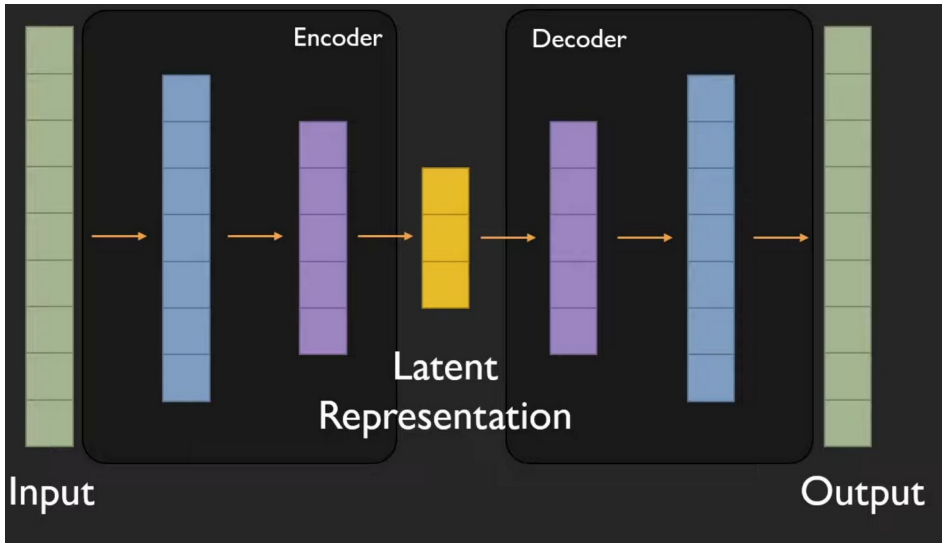
Les CNN utilisent des représentations hiérarchiques

Masked Autoencoder (MAE)

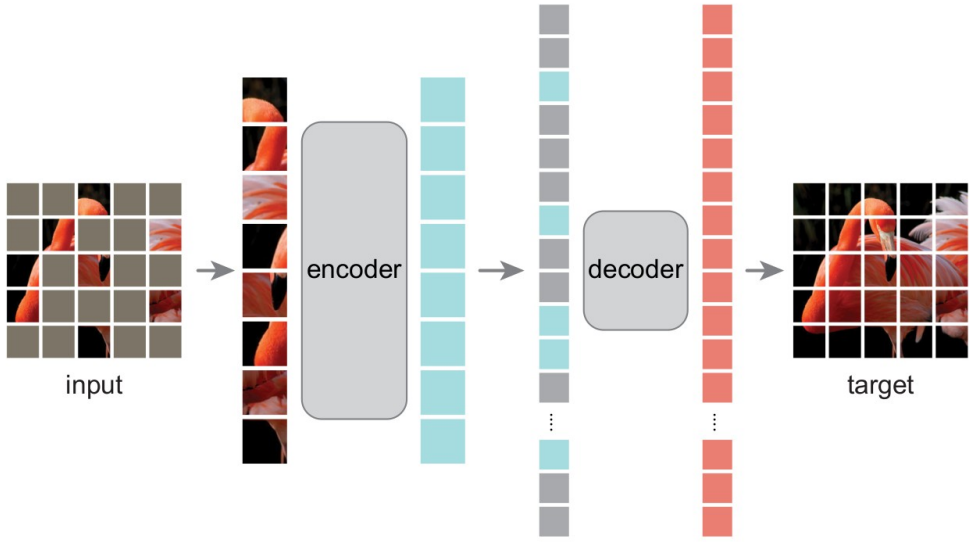
MAE



Autoencoder



Masked Autoencoder



source

Fonctionne en auto-supervision.

Masked Autoencoder (MAE)

ViT trained as MAE

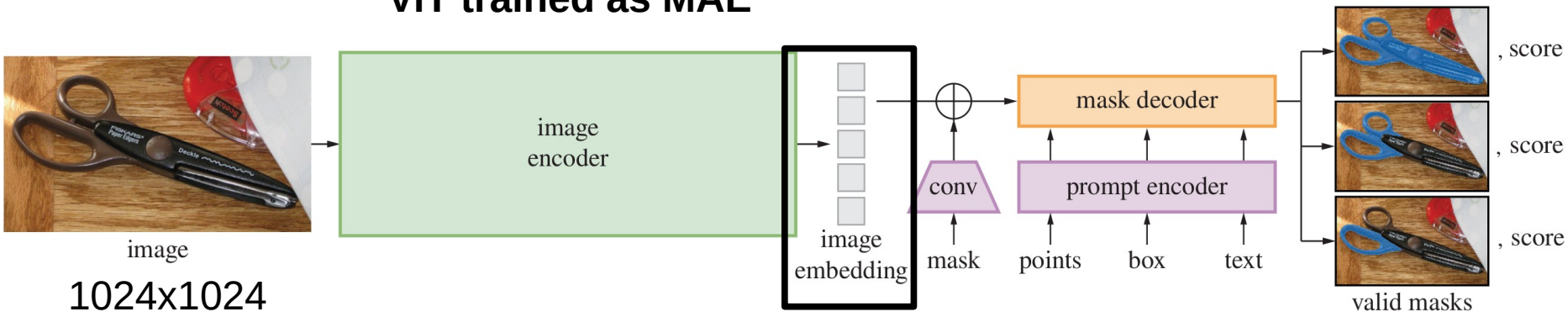
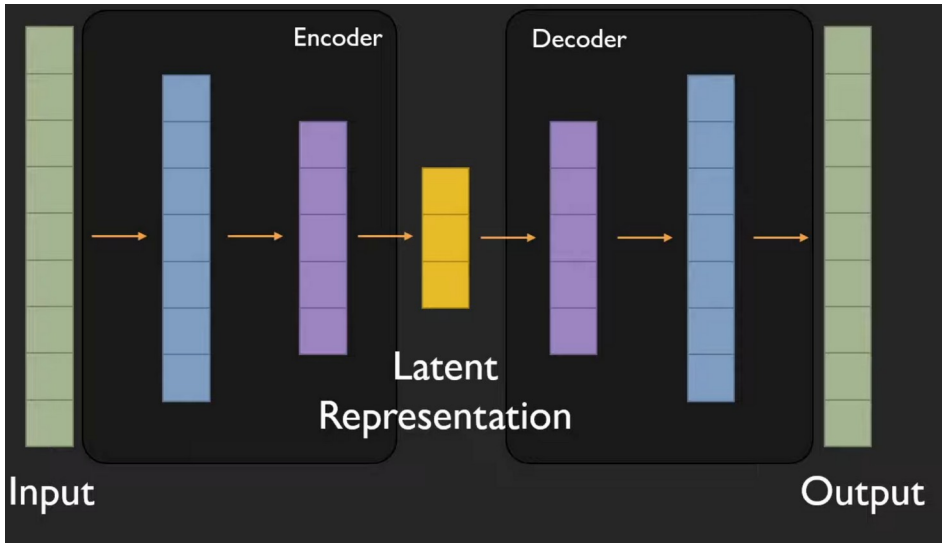


image
1024x1024

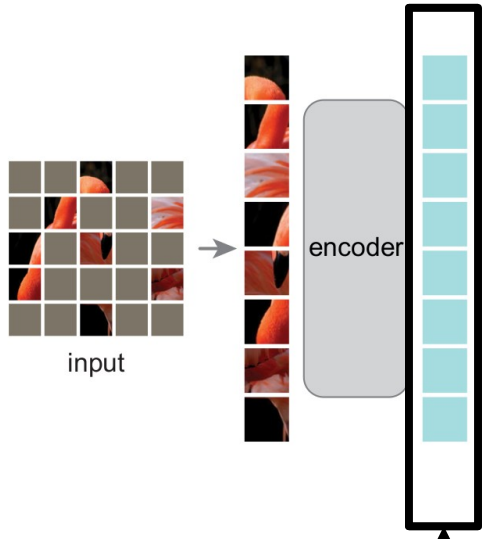
256x64x64

(#tokens)x256

Autoencoder

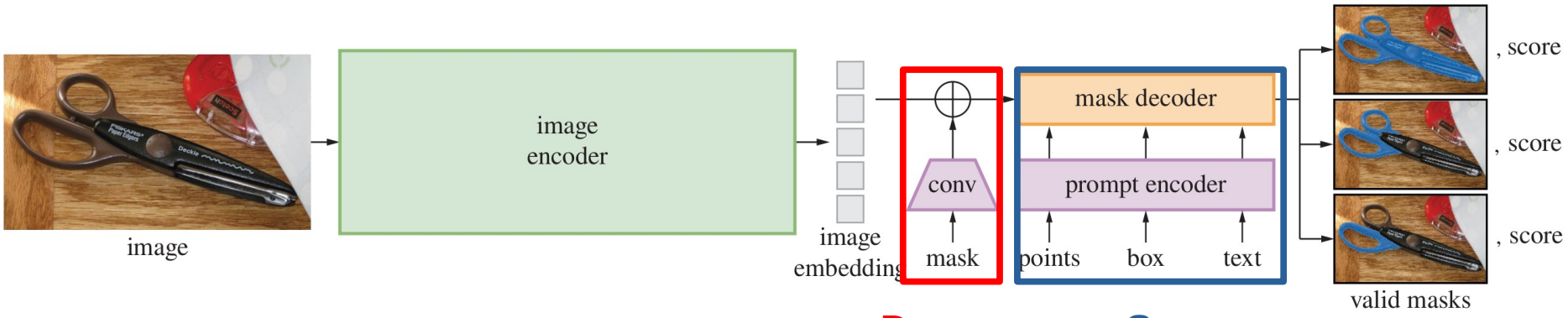


Masked Autoencoder



représentations apprises

Segment Anything

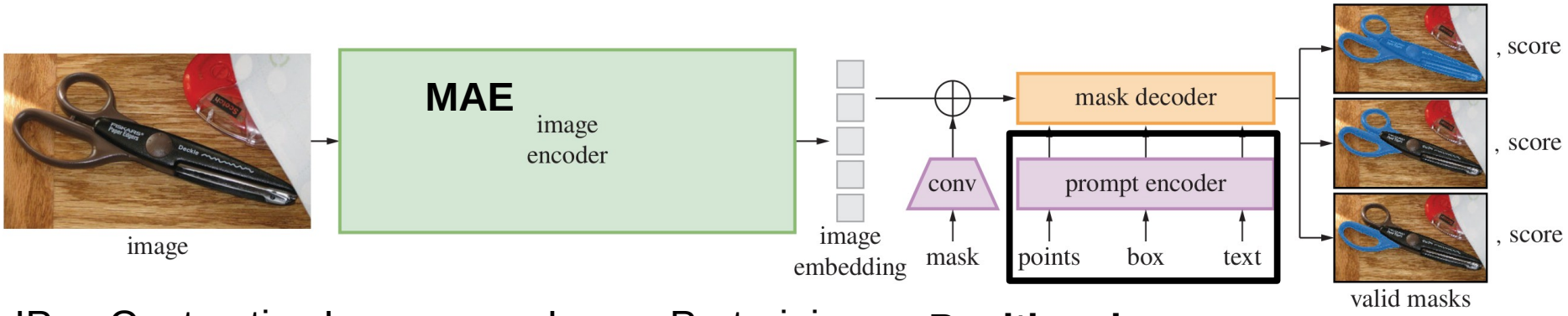


Dense prompts

Sparse prompts

50 ms temps d'inférence dans un navigateur web

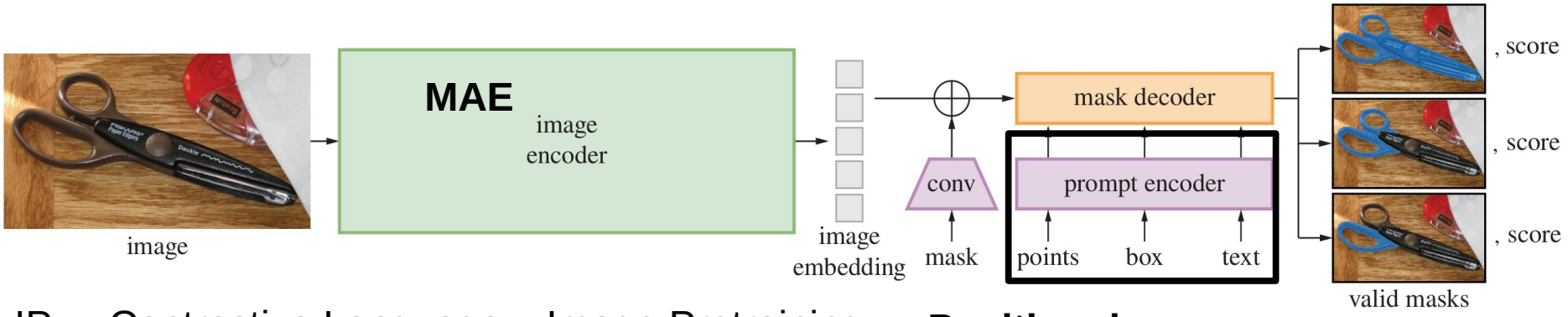
Segment Anything



CLIP = Contrastive Language – Image Pretraining

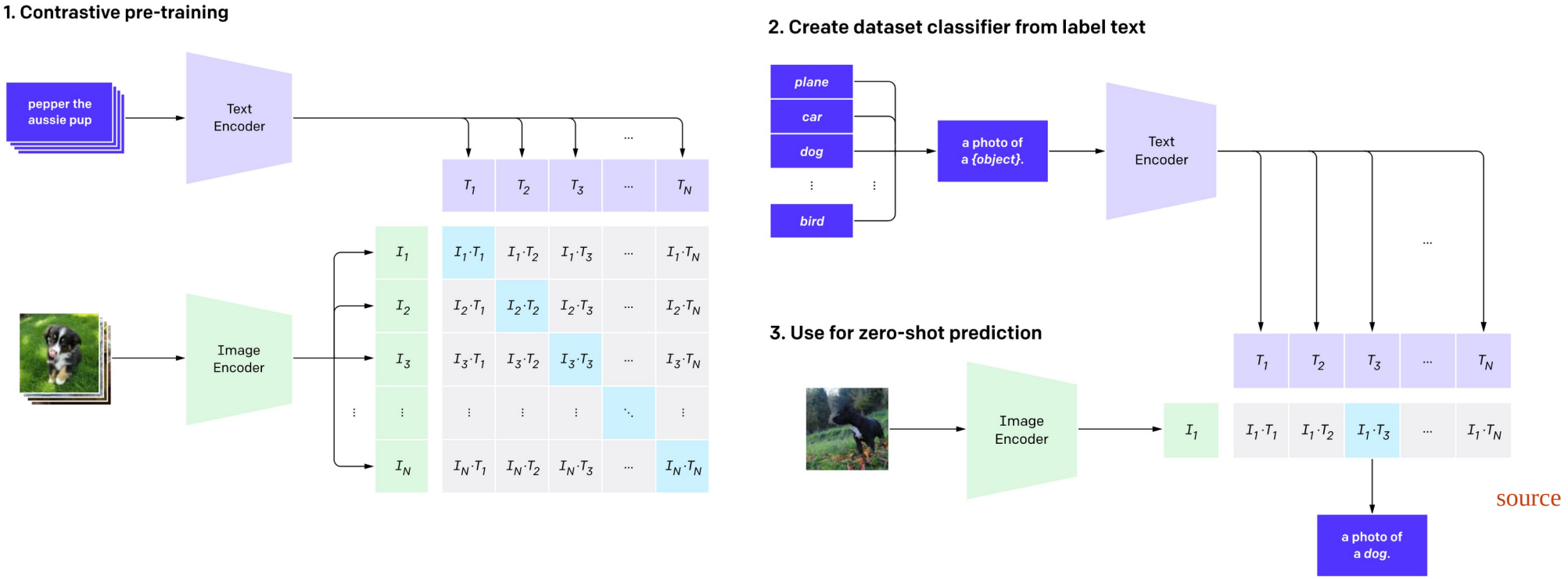
Positional encoddings **+ CLIP model**

Segment Anything



CLIP = Contrastive Language – Image Pretraining

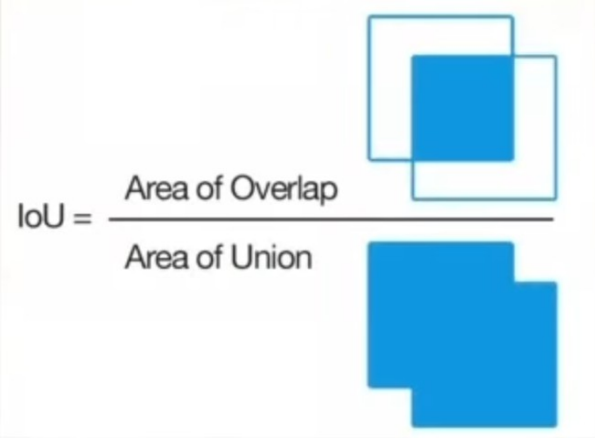
Positional encoddings **+ CLIP model**



source

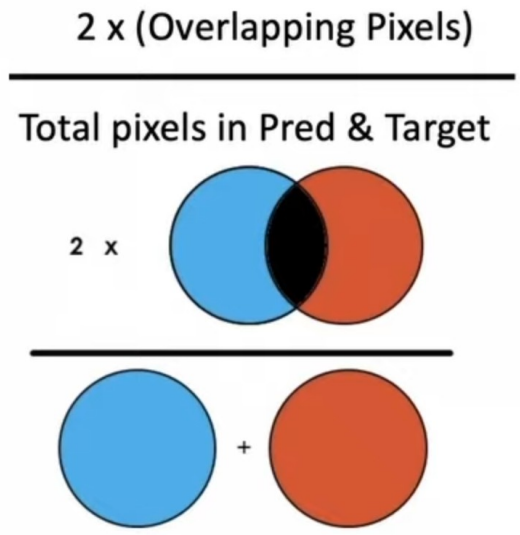
L'entraînement

Pour chaque objet – produire 3 masques : entier, partie, sous-partie



les masques ayant le plus haut IoU avec les données de référence est utilisé pour calculer la perte (loss)

DICE LOSS **+** **FOCAL LOSS**



$$FL = -(1 - P_t)^{\gamma} \log(P_t)$$

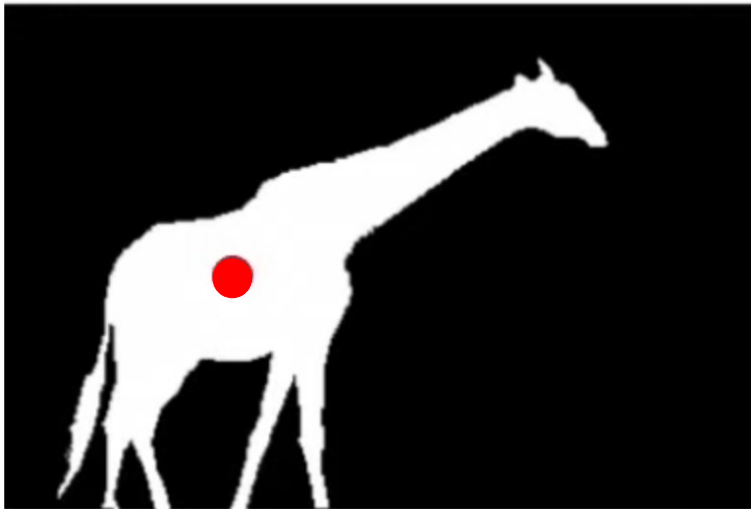
$$CE = -\log(P_t)$$

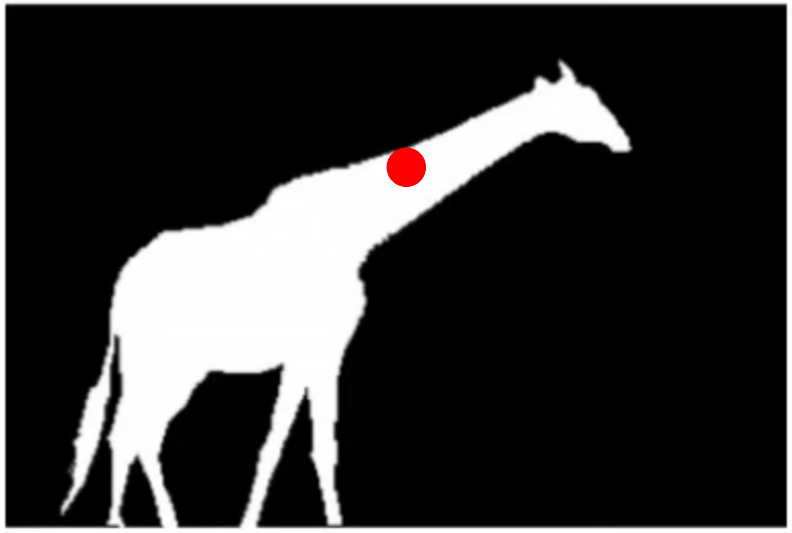
L'entraînement interactif

image d'entraînement

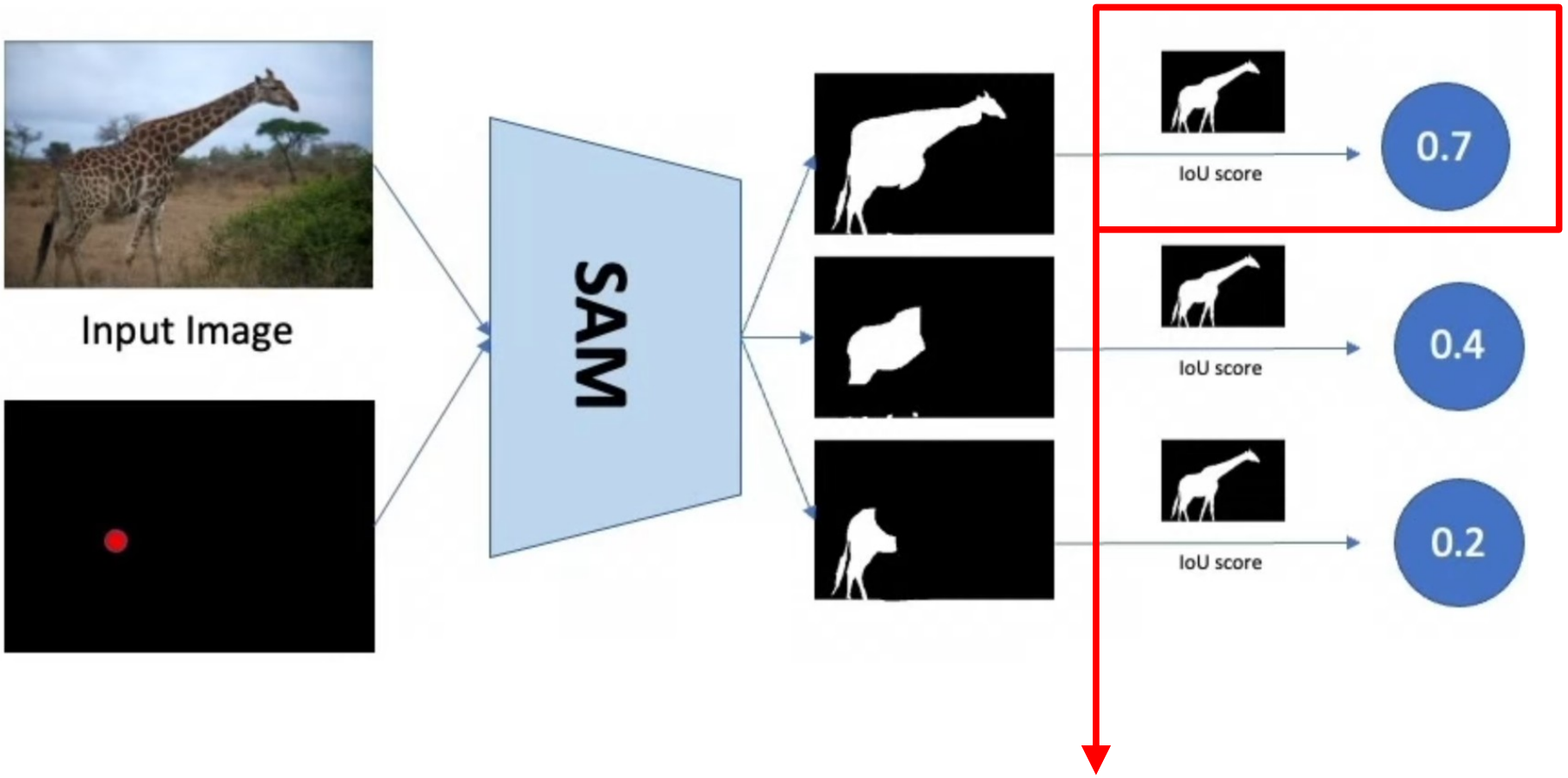
image cible

Prompt aléatoire





L'entraînement interactif



L'entraînement interactif

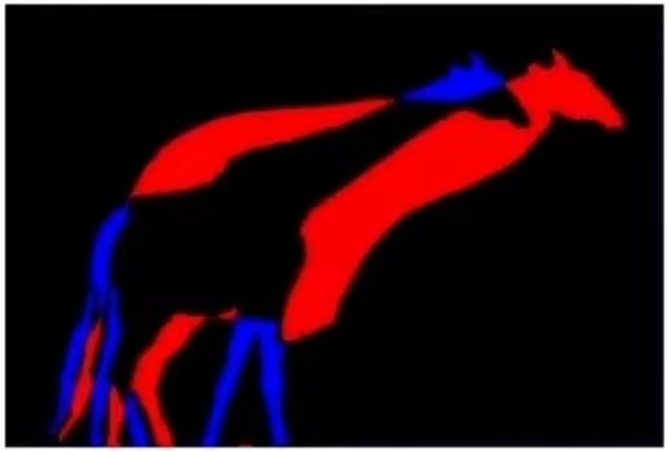
image cible | prédiction | région d'erreur

la région d'erreur = la différence entre la cible et le masque prédit

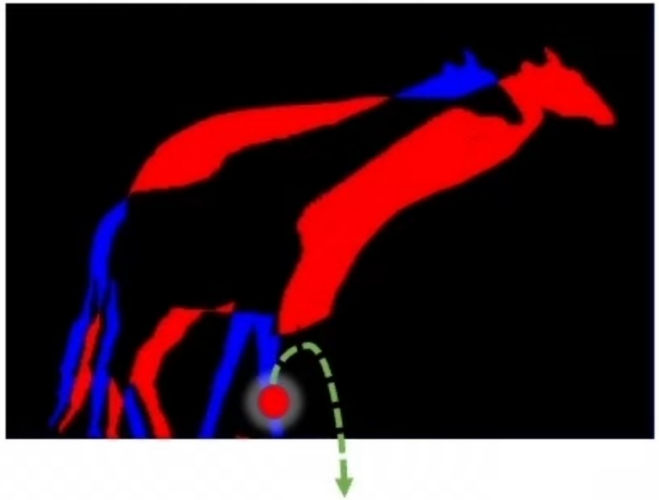
L'entraînement interactif

image cible

prédiction

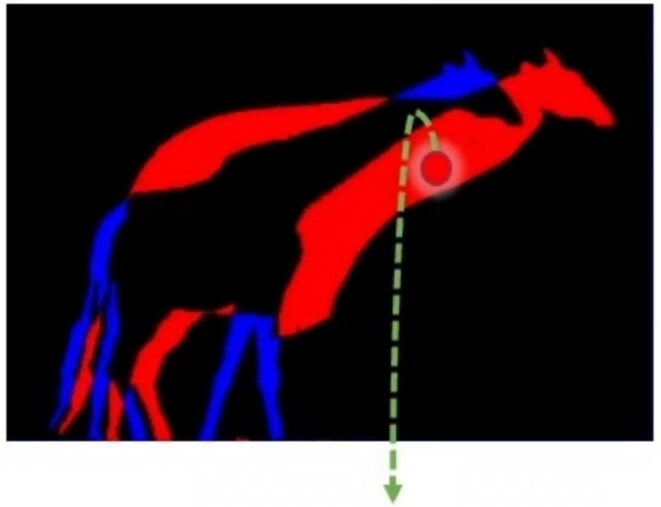


région d'erreur



Faux négatif

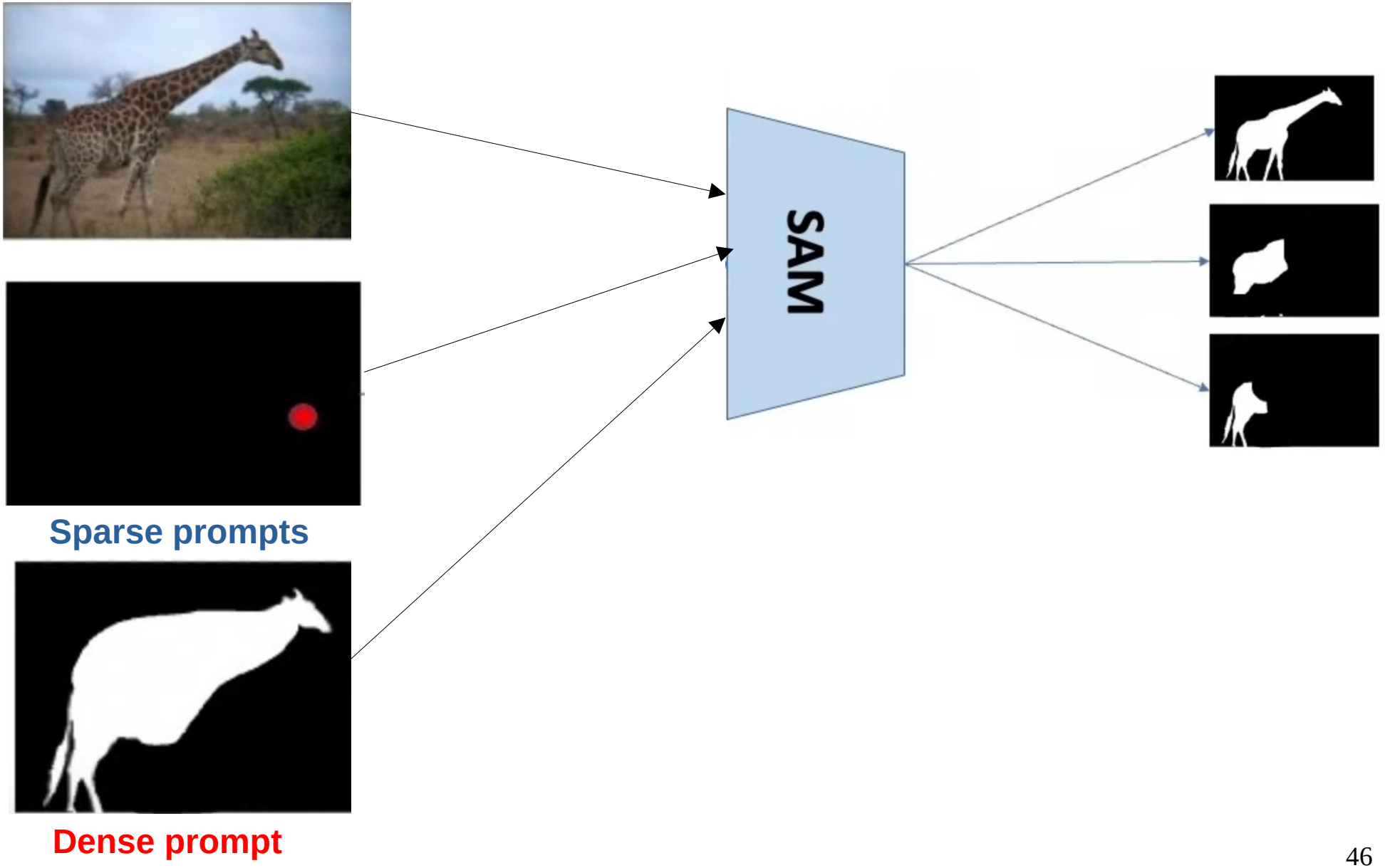
Point d'avant-plan
(foreground)



Faux positif

Point d'arrière-plan
(background)

L'entraînement interactif



Segment Anything

DEMO