

Applications Avancées des CNN en Vision par Ordinateur

Détection d'objets, Segmentation et Transformers

Florian Valade

Université Gustave Eiffel

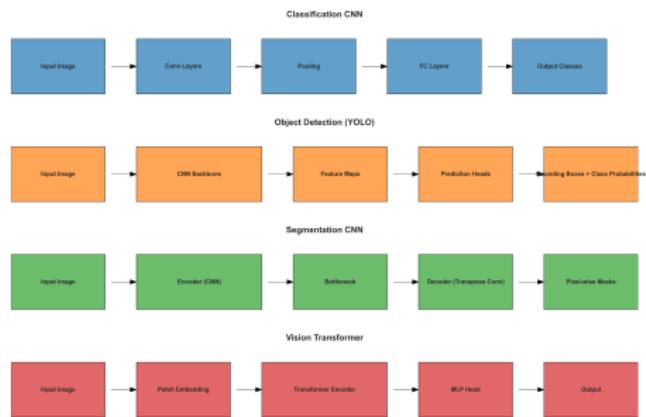
13 mars 2025

Plan du cours

- 1 Introduction
- 2 Détection d'objets
- 3 Segmentation
- 4 Transformers en Vision
- 5 SAM (Segment Anything Model)
- 6 Conclusion

Rappel des fondamentaux

- Nous avons déjà vu :
 - Les réseaux de neurones (MLP)
 - Les couches de convolution
 - Les CNN pour la classification
- Aujourd'hui : applications avancées



Au-delà de la classification

- Classification : "*Qu'y a-t-il dans l'image ?*"
- Détection : "*Où sont les objets et que sont-ils ?*"
- Segmentation : "*Quels pixels appartiennent à chaque objet ?*"
- Autres tâches :
 - Pose estimation
 - Instance segmentation
 - Panoptic segmentation
 - Depth estimation

Évolution des tâches

Classification



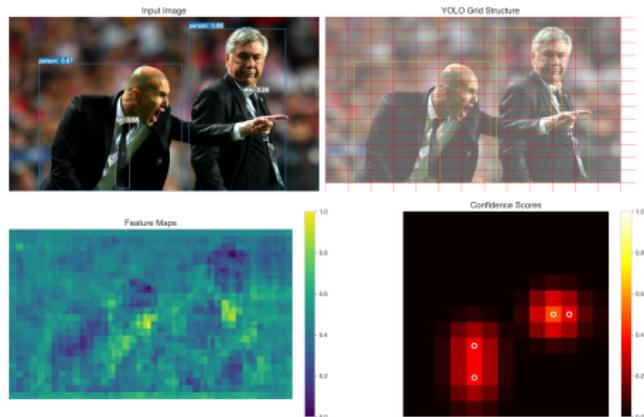
Détection



Segmentation

Détection d'objets : Principes

- Classification + Localisation
- Sortie : boîtes englobantes (bounding boxes) + classes
- Mesures d'évaluation :
 - IoU (Intersection over Union)
 - mAP (mean Average Precision)
- Challenges :
 - Objets de tailles différentes
 - Objets multiples
 - Chevauchement d'objets

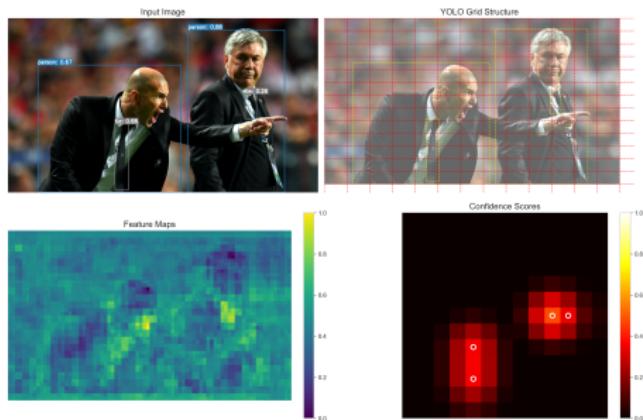


Approches de détection

- Approches en deux étapes :
 - R-CNN, Fast R-CNN, Faster R-CNN
 - Génère des propositions de régions, puis classification
 - Plus précis mais plus lent
- Approches en une étape :
 - YOLO, SSD, RetinaNet
 - Prédiction directe des boîtes et classes
 - Plus rapide, adapté au temps réel

YOLO : You Only Look Once

- Approche révolutionnaire (2016)
- Divise l'image en grille
- Chaque cellule prédit :
 - Boîtes englobantes
 - Scores de confiance
 - Probabilités de classes
- YOLOv1, v2, v3, v4, v5, v7...
- Équilibre vitesse/précision



Architecture de YOLO

- Backbone : extrait les caractéristiques (ex : Darknet)
- Neck : fusionne les caractéristiques de différentes échelles (FPN, PANet)
- Head : prédit les boîtes et les classes

Prédiction par cellule

Chaque cellule prédit :

- B boîtes (x, y, w, h)
- Score de confiance pour chaque boîte
- C probabilités de classes

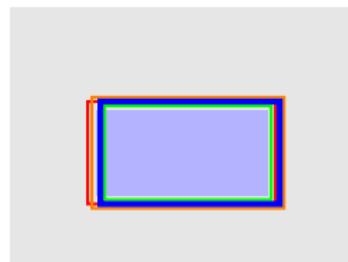
Total : $B \times 5 + C$ prédictions par cellule

Avantages de YOLO

- Rapide (45-155 FPS)
- Raisonnement global
- Généralisation
- Apprentissage end-to-end

Non-Maximum Suppression (NMS)

- Problème : détections multiples du même objet
- Solution : Non-Maximum Suppression
- Algorithme :
 - ① Trier les boîtes par score de confiance
 - ② Sélectionner la boîte avec le score le plus élevé
 - ③ Supprimer les boîtes avec $\text{IoU} > \text{seuil}$
 - ④ Répéter jusqu'à ce qu'il n'y ait plus de boîtes



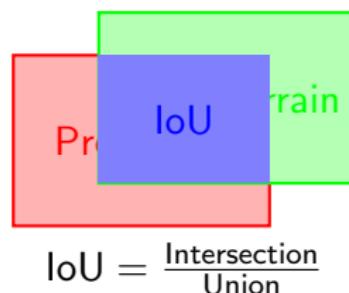
Avant et après NMS

Métriques d'évaluation

- IoU (Intersection over Union) :

$$\text{IoU} = \frac{\text{Aire d'intersection}}{\text{Aire d'union}}$$

- Precision : $\frac{TP}{TP+FP}$
- Recall : $\frac{TP}{TP+FN}$
- Average Precision (AP) : aire sous la courbe Precision-Recall
- mAP : moyenne des AP sur toutes les classes
- AP@0.5, AP@0.75, AP@[.5 :.95]



① Prédictions multiples

- Le modèle prédit de nombreuses boîtes à chaque inférence
- Plusieurs prédictions peuvent recouvrir le même objet
- Des scores de confiance sont associés à chaque prédition

② Filtrage des prédictions

- **Seuil de confiance** : Élimination des boîtes avec score < seuil
- **Non-Maximum Suppression (NMS)** : Élimination des doublons
 - On garde la boîte avec le meilleur score
 - On supprime les boîtes avec $\text{IoU} > \text{seuil NMS}$

Entraînement des détecteurs (2/2) : Calcul de la loss

① Association prédition-vérité terrain

- Chaque prédition est associée à une annotation si IoU > seuil (souvent 0.5)
- Cette association détermine les cibles positives et négatives

② Composantes de la fonction de perte

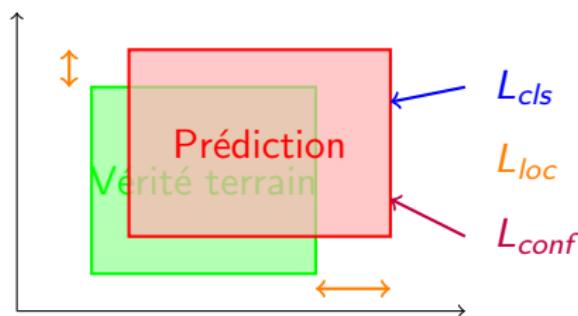
- **Loss de localisation (L_{loc})** : Écart entre coordonnées prédites et réelles
 - Smooth-L1 ou MSE sur (x, y, w, h) des boîtes
- **Loss de classification (L_{cls})** : Erreur sur les classes prédites
 - Cross-entropy ou Focal Loss
- **Loss de confiance (L_{conf})** : Objectness (présence d'objet)
 - Binary cross-entropy sur le score de confiance

③ Loss totale : $L = \lambda_{loc} L_{loc} + \lambda_{cls} L_{cls} + \lambda_{conf} L_{conf}$

- λ sont des hyperparamètres de pondération

Visualisation du calcul de la loss

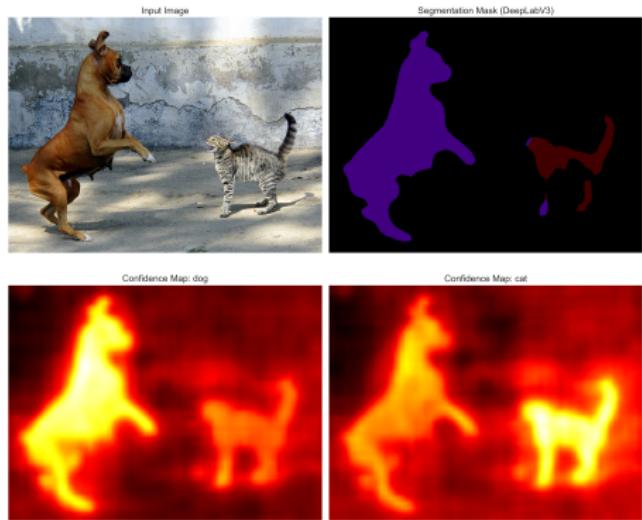
Calcul de la loss



$\text{IoU} = 0.62 > 0.5 \Rightarrow \text{Association positive}$

Introduction à la segmentation

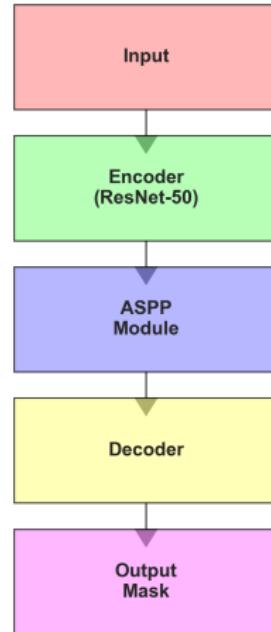
- Classification au niveau des pixels
- Types de segmentation :
 - Sémantique : classe pour chaque pixel
 - Instance : sépare les objets de même classe
 - Panoptique : combine sémantique et instance
- Applications :
 - Voitures autonomes
 - Imagerie médicale
 - Réalité augmentée



Architectures pour la segmentation

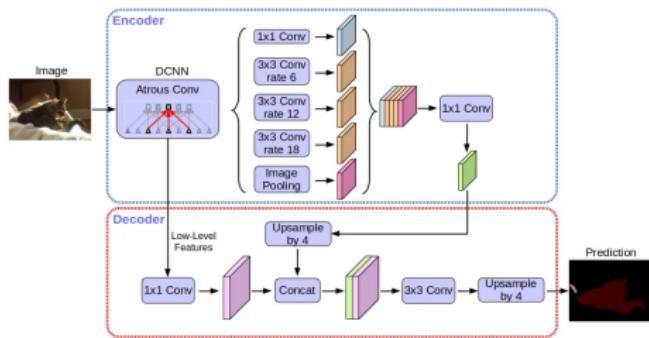
DeepLabV3 Architecture

- Architecture encodeur-décodeur
- Encodeur : réduit la dimension spatiale, extrait les caractéristiques
- Décodeur : restaure la dimension spatiale, affine les prédictions
- Skip connections : préservent les détails spatiaux
- Exemples :
 - FCN (Fully Convolutional Networks)
 - U-Net
 - DeepLab
 - Mask R-CNN



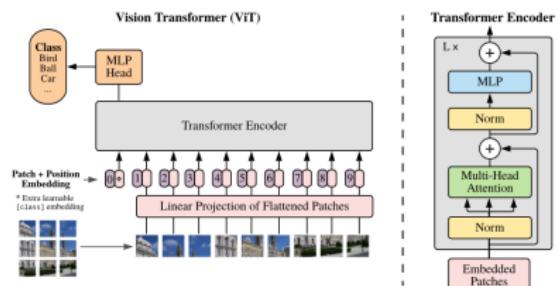
DeepLab

- Série d'architectures DeepLabv1, v2, v3, v3+
- Innovations clés :
 - Convolutions dilatées
 - ASPP (Atrous Spatial Pyramid Pooling)
- Avantages :
 - Champ réceptif large
 - Multi-échelle
 - Préservation de la résolution



Des Transformers au ViT

- Transformers : révolution en NLP (2017)
- Vision Transformer (ViT, 2020) :
 - Divise l'image en patches
 - Traite les patches comme des tokens
 - Utilise le mécanisme d'attention
- Avantages :
 - Capture des dépendances à longue distance
 - Parallélisable (vs. CNN séquentiels)
 - Performances impressionnantes



Mécanisme d'attention

- Mécanisme central des Transformers
- Auto-attention : permet à chaque position d'interagir avec toutes les autres
- Calcul d'attention :

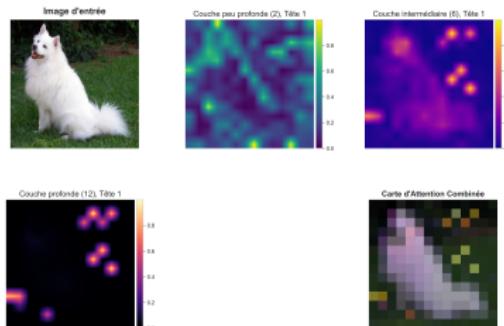
$$\text{Attention}(Q, K, V)$$

$$= \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \cdot V$$

où :

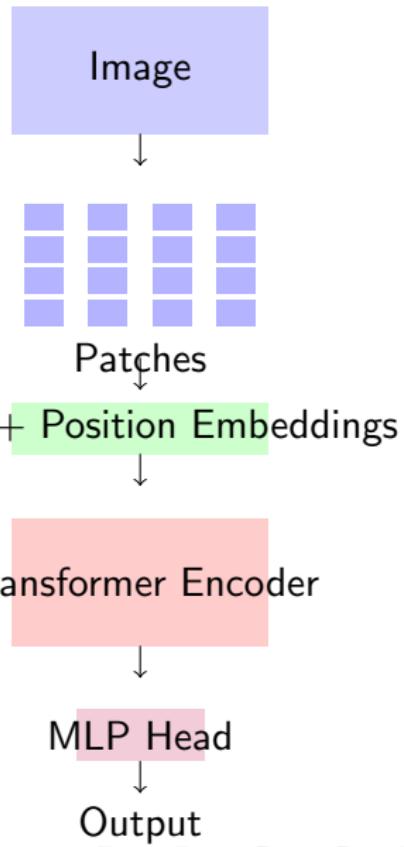
- Q : requêtes (queries)
- K : clés (keys)
- V : valeurs (values)
- d_k : dimension des clés

Mécanisme d'attention dans les Vision Transformers



Architecture ViT

- ➊ Division de l'image en patches (ex : 16x16)
- ➋ Projection linéaire des patches
- ➌ Ajout d'embeddings de position
- ➍ Séquence de blocs Transformer :
 - Multi-Head Attention
 - MLP (Multi-Layer Perceptron)
 - Layer Norm et résidus
- ➎ MLP final pour la classification



Modèles basés sur les Transformers

- DETR (DEtection TRansformer)
 - Détection d'objets sans NMS ni ancrès
 - Prédit directement un ensemble de boîtes
- Swin Transformer
 - Fenêtres d'attention glissantes
 - Hiérarchique, comme les CNN
- DeiT (Data-efficient image Transformer)
 - Entraînement efficace avec moins de données
 - Distillation de connaissances
- SegFormer, Mask2Former...
 - Transformers pour la segmentation

Introduction à SAM

- Modèle de Meta AI (2023)
- Premier Foundation Model pour la segmentation
- Caractéristiques :
 - Pré-entraîné sur 11M d'images, 1B de masques
 - Zero-shot : fonctionne sur de nouvelles tâches
 - Prompt-based : points, boîtes, texte
 - Temps réel



Architecture de SAM

- Architecture en trois parties :

- ① Encodeur d'image

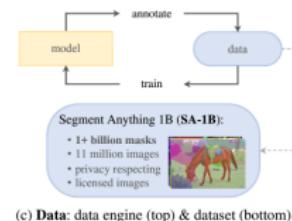
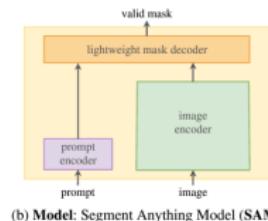
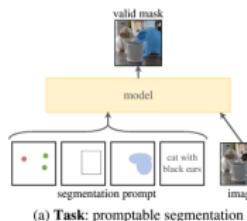
- MAE-ViT pré-entraîné
 - Génère des embeddings d'image

- ② Encodeur de prompts

- Encode points, boîtes, texte

- ③ Décodeur de masque

- Transformer léger
 - Prédit des masques



Utilisation de SAM

- Modes d'utilisation :
 - Automatique : segmentation sans prompt
 - Interactive : points/boîtes par l'utilisateur
 - Textuel : description de l'objet à segmenter

Forces de SAM

- Zéro-shot
- Adaptabilité
- Interactivité
- Précision des contours
- Masques multiples
- Temps réel

Récapitulatif

- Nous avons vu :
 - La détection d'objets avec YOLO
 - La segmentation (U-Net, DeepLab, Mask R-CNN)
 - Les Transformers en vision (ViT)
 - SAM, premier foundation model pour la segmentation
- Évolution des architectures :
 - CNNs → Transformers → Modèles hybrides
 - Modèles spécialisés → Foundation models

Tendances futures

- Modèles multimodaux
- Apprentissage auto-supervisé
- Modèles plus légers (mobile, edge)
- IA générative pour la vision
- Vision 3D et scènes complexes

Merci !

Des questions ?