Unleashing the Potential of ConvNets for Query-Based Detection and Segmentation

Présenté par Hatem Ayed Zhile Zhang Salwa Mujahid

Sorbonne Université

Introduction

Les modèles basés sur les Transformers comme DETR ont permis des avancées majeures en détection d'objets, mais leur coût computationnel élevé limite leur déploiement. Cet article propose InterConv, un mécanisme qui imite l'interaction entre les requêtes d'objets et les caractéristiques de l'image, habituellement traitée par les mécanismes d'attention dans les Transformers, en utilisant uniquement des couches convolutionnelles. L'objectif est de montrer que les réseaux convolutifs peuvent rivaliser avec les Transformers en préservant des performances élevées en détection et segmentation tout en améliorant l'efficacité computationnelle.

Architecture du Modèle DECO (Detection ConvNet)

- **Encodeur**: Basé sur ConvNeXt avec kernel 7×7 , extraction des caractéristiques spatiales riches sans nécessiter d'encodage positionnel explicite.
- **Décodeur basé sur InterConv**: Remplacement de l'attention par des convolutions optimisées.
  - **Self-Interaction Module (SIM)**: Interaction entre les requêtes via une convolution en profondeur 9×9 appliquée sur une grille restructurée, il permet de capturer les relations locales sans dépendre d'une attention globale.
  - **Cross-Interaction Module (CIM)**: Fusion des requêtes avec les features visuelles en utilisant un échantillonnage et une convolution en profondeur:

1. Upsampling des object queries:

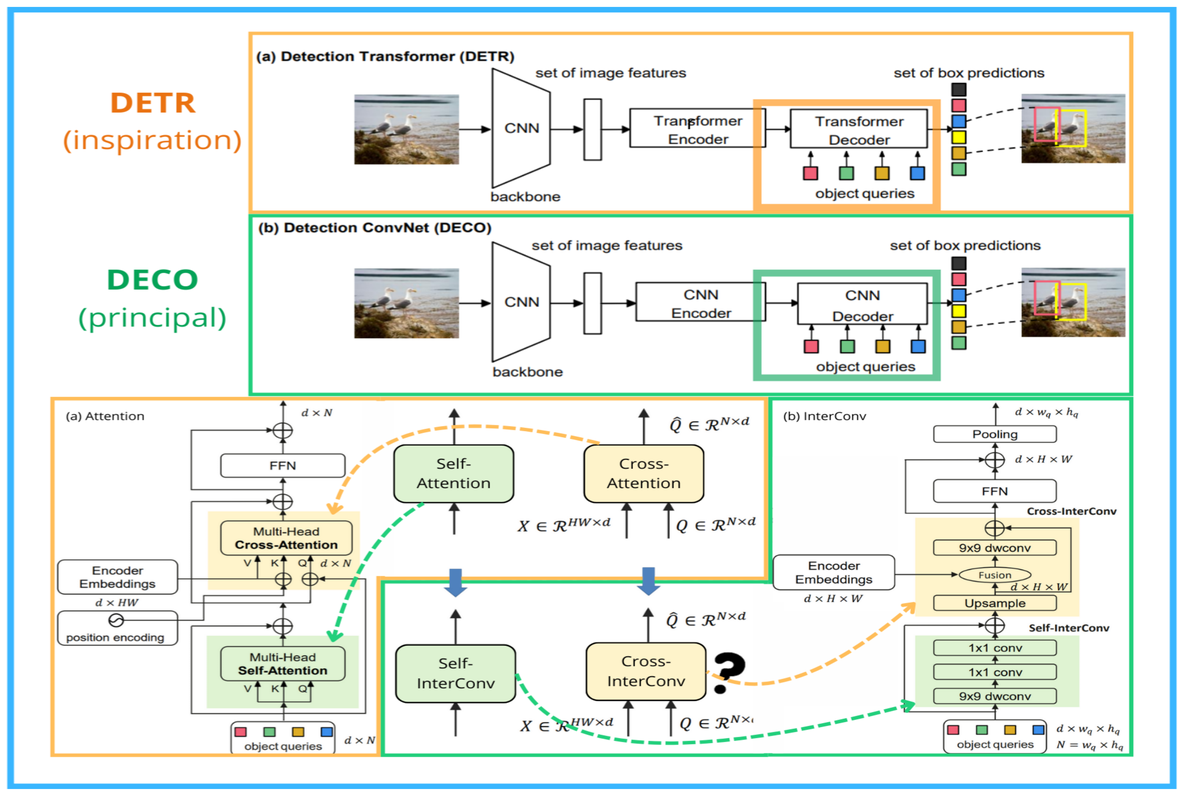
$$\hat{o} = \text{Upsample}(o) \tag{1}$$

2. Fusion des upsampled object queries et les image feature embeddings avec F (Element-wise Add):

$$F = \text{Fusion}(\hat{o}, Z_e) \tag{2}$$

3. Extraction d'information spatiale (grands noyaux):

$$F' = \text{dwconv}(F) \tag{3}$$



Implémentation

Architecture:

- Backbone : ResNet18 pré-entraîné sur ImageNet
- Decoder : 6 couches avec Self-Interaction Module (SIM) et Cross-Interaction Module (CIM)

Configuration:

- Queries: Grille 10×10 (100 queries) (optimal, cf Ablation study)
- Convolutions: Kernel 9×9 pour SIM/CIM

Dataset: COCO. (Augmentation: Random Horizontal Flip, Random Resize, Random Crop)

Discussion

InterConv est une approche innovante permettant l'interaction entre les object queries et les image feature embeddings via des couches convolutionnelles, remplaçant ainsi l'architecture Transformer.
DECO démontre qu'un décodeur convolutif peut rivaliser avec les Transformers, mais son adaptation à d'autres tâches et datasets doit être explorée pour confirmer sa robustesse.
De plus, des études d'ablation supplémentaires pourraient être menées pour évaluer l'impact individuel de chaque composant sur les performances globales.

Résultats

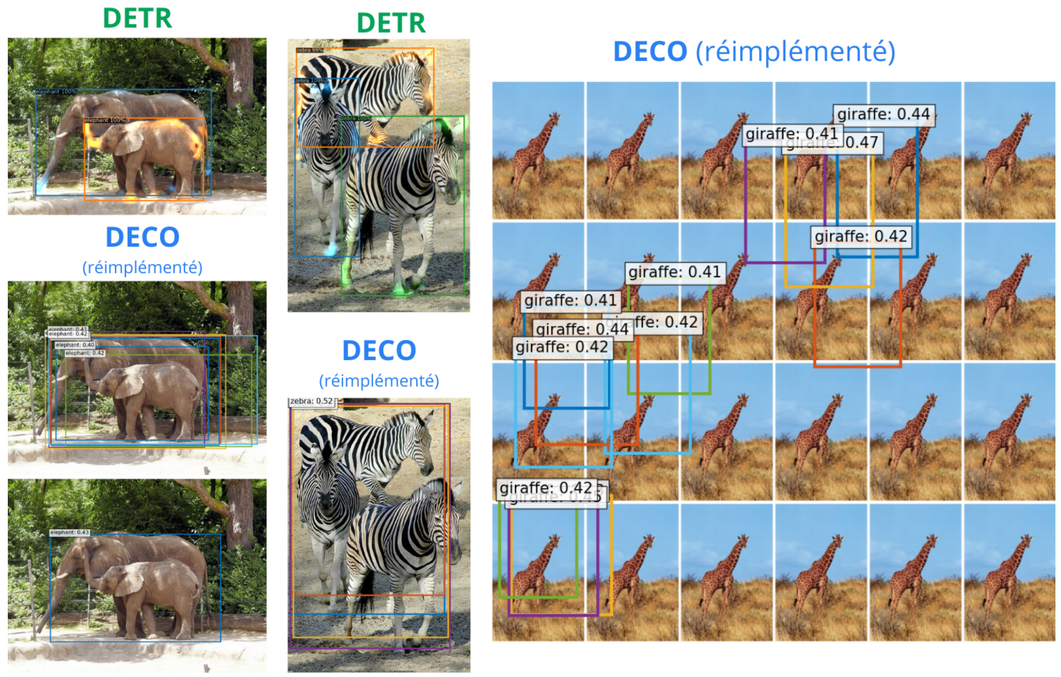


Figure 1:Comparaison des résultats de détection du DECO réimplémenté et DETR

Modèle	AP(Average Precision)			AR(Average Recall)		
	Small	Medium	Large	Small	Medium	Large
DETR(Carion et al., 2020)	0.175	0.43	0.591	null	null	null
DECO (papier) R50	0.195	0.434	0.55	null	null	null
DECO (ours) R18	0.005	0.027	0.103	0.021	0.159	0.456

Table 1:Comparaison des performances du modèle DECO (papier vs ours) et du modèle DETR