Rapport du support avancé “Object-centric representation for robotic manipulation”

- **Nom du code :** Segment-Obj-Centric
- **Partenaire :** CHAPIN Alexandre
- **Personnel IDRIS :** SONG Maxime
- **Dates et durée du support :** Septembre 2023 - Mars 2024

1 Description du code

Basé sur le papier https://arxiv.org/pdf/2206.06922, un premier modèle a été entraîné de manière auto-supervisé pour extraire, depuis des images naturelles, des représentations vectorielles des objets de la scène, cette extraction utilisait un composant spécialement conçu appelé Slot Attention qui prenait en entrée plusieurs représentation vectorielle de la scène (on avait plusieurs prises photos de la même scène). En reprenant ce modèle, plusieurs pistes d’améliorations du modèle ont été envisagées.

2 Travail effectué et résultats obtenus

Objectif : Améliorer l’extraction et la représentation vectorielle des objets de la scène, afin de pouvoir plus tard le ré-utiliser pour des robots.

Travail effectué : Un premier travail a été de préparer l’entraînement du modèle de base en multi-gpus sur Jean Zay, le scaling allant jusqu’à 8 GPUs avec un gain en entraînement presque de x8. Ensuite on a envisagé d’augmenter le réseau de neurones existant en ajoutant à la Slot Attention, l’extraction des masques de segmentation avec Segment Anything Model (SAM) de Meta. On a lancé des expériences sur Jean Zay avec les poids de SAM freezé, mais cette approche n’a pas été concluante, les résultats étant moins bons que le modèle initial, notamment parce qu’on se retrouvait souvent avec des masques de segmentations vides. Pour pallier cela, on a entraîné SAM avec le modèle, mais on a eu des problèmes techniques avec le positional encoding dans SAM, étant donné qu’on travaillait sur des plus petites images (240x320) qu’attendues par SAM (1024x1024).

Ensuite, on a été intéressé par le modèle Dino V1. Celui-ci repose sur l’architecture Vision Transformer et un entraînement auto-supervisé pour faire émerger naturellement dans ses cartes d’attentions, les segmentations des objets. Des scripts et un benchmarking ont été réalisés sur Dino V1.