Segmentation interactive d'images
modèle SAM appliqué à des photos de poissons

Ana Bernal[1]
Mentor: Samir Tanfous[2]
avril 2023

1. Introduction

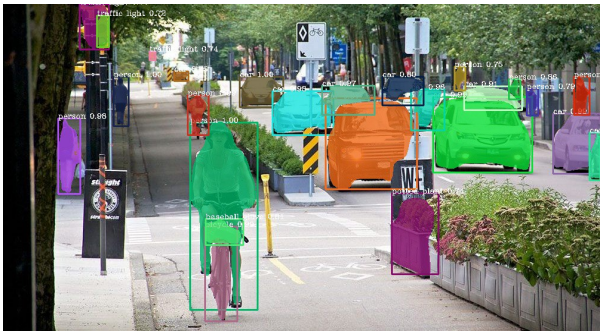
Ceci est un rapport sur le projet d'application du modèle SAM. Le modèle SAM est un modèle qui vient d'être publié (voir [1]) par des chercheurs de Meta AI. C'est un modèle de segmentation d'image interactif : il suffit d'un clic de l'utilisateur pour sélectionner un objet sur une image.

Pour ce projet, j'ai étudié l'article où ce modèle a été introduit. Pour cela, j'ai dû me contextualiser dans le domaine de la segmentation d'image avec des réseaux de neurones. Ensuite j'ai cherché un ensemble des données différentes à celles évaluées dans l'article, et sur lesquelles j'ai appliqué l'algorithme. Pour évaluer les performances, j'ai cherché et exerimenté avec un modèle précédent comme baseline.

Ce rapport est un bref résumé de toute cette procédure.

2. Segmentation d'image : quelques définitions et état de l'art

Cette section contient un court résumé et notions sur la segmentation d'image à l'aide des réseaux de neurones. Elle est basée en partie dans l'article [2].



La segmentation d'image est une tâche clé dans le domaine de traitement d'image et vision par ordinateur. Cela consiste à sélectionner, distinguer et séparer des objets ou instances dans une image, voir l'image suivante.
La tâche de segmentation a beaucoup d'applications importantes et très utiles, comme l'analyse d'images médicales, la perception robotique, la vidéosurveillance, la réalité augmentée et la compression d'images, entre autres.
Le succès et performances importantes obtenus grâce à l'apprentissage automatique (machine learning) avec des réseaux de neurones profonds a permis une grande avancée dans cette tâche.

3. Définitions et traduction du problème

3.1. Segmentation d'image : C'est un problème de classification.

Les algorithmes de réseaux de neurones prennent en entrée des matrices. Dans le problème de segmenter une image, nous partons d'une image ! c'est-à-dire d'un fichier `.jpeg`, par exemple. Nous devons donc traduire l'image en matrice. Cela se fait en associant chaque pixel aux trois nombres correspondants à la couleur en code RGB.
Une fois que l'image devient un array: une matrice tridimensionnelle de nombres, le problème de segmentation d'image devient **un problème de classification** : si nous cherchons par exemple à segmenter un chien dans l'image, nous cherchons à classifier chaque pixel de l'image dans le deux catégories : chien ou non-chien. Plus généralement, si nous voulons segmenter toutes les instances qui apparaissent dans l'image cela peut être vu comme un problème de classification multiclasse, où

[1] E-mail : anabeatrizbernal@gmail.com
[2] Mentor OpenClassrooms

chaque pixel doit être classifié dans une classe, par exemple: chien, arbre, herbe, table, réfrigérateur, etc. Ce type de segmentation est appelée **segmentation sémantique.**

3.2. État de l'art

Concrètement, les dernières avancées dans la segmentation d'images ont été faites avec des algorithmes de réseaux de neurones profondes, avec différents types d'architecture, parmi lesquelles on peut retrouver : Réseaux entièrement convolutifs, modèles basés sur un codeur-décodeur, modèles basés sur des réseaux multi-échelles et pyramidaux, modèles basés sur l'attention, etc.
Ces structures sont assez techniques pour être expliquées en détail dans ce rapport. Voici une sélection de deux idées, termes clés et références pour plus de détails.

- **Réseaux de neurones convolutifs** (**CNN**, pour les sigles en anglais). Les CNN sont parmi les architectures les plus réussies et les plus utilisées dans la communauté de l'apprentissage profond, en particulier pour les tâches de vision par ordinateur. Les CNN ont été initialement proposés par Fukushima dans les années 80 dans son article fondateur sur le "Neocognitron" (voir [3]). L'idée intuitive est qu'une couche de convolution prend en entrée une image et elle va apprendre les caractéristiques distinctives de l'objet qu'illustre l'image, sans que l'on doit le spécifier. Ces caractéristiques sont enregistrées en tant que "cartes de caractéristiques" ou *feature maps* en anglais. Voici un diagramme illustratif de l'architecture.

- **Modèles basés dans l'Attention.** Chen et al [4] ont proposé un mécanisme d'attention qui apprend à pondérer doucement les caractéristiques multi-échelles à chaque pixel.
  Ils adaptent un puissant modèle de segmentation sémantique et l'entraînent conjointement avec des images multi-échelles et le modèle d'attention.

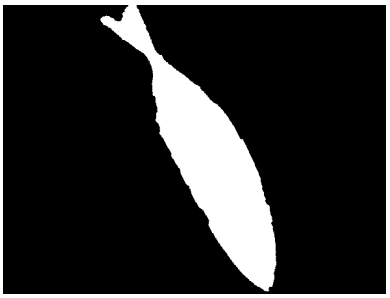
3.3. Métriques : mesure des performances.

Puisque ce n'est pas l'objectif de ce rapport de parler en détail de ces détails techniques, nous pouvons aborder un sujet important et c'est celui de l'évaluation : comment évaluer un modèle de segmentation d'image ?

Avant de parler de ces métriques, on introduit le terme **carte de segmentation**, ou **mask**, en anglais. La carte de segmentation est, dans notre cas, une matrice avec des 0 et des 1 (ou des booléens) qui distinguent la localisation de l'objet à segmenter dans l'image. Nos ensembles de données consistent donc à des images et à ses masks qui sont les appelées **ground truth.** Nous utiliserons ce terme pour parler de la mask d'origine qui permet de distinguer l'objet.

Le modèle de segmentation va, lui, produire des masks et c'est avec les masks de ground truth que l'on va les comparer pour évaluer les performances. Les masks de ground truth sont donc les **labels** d'origine.

L'image



Sa ground truth mask

Il y a plusieurs méthodes pour évaluer les performances d'un modèle de segmentation. Quelques unes sont:

- **Pixel accuracy :** c'est le rapport de pixels correctement classifiés divisé par le nombre total de pixels. La Mean Pixel Accuracy (MPA) fait la moyennes de ces valeurs par classe d'objet.
- **Intersection over Union** (**IoU**) ou l'**indice de Jaccard** est l'une des mesures les plus couramment utilisées dans la segmentation sémantique. Il est défini comme la zone d'intersection entre la carte de segmentation (mask) prédite et la ground truth, divisée par la zone d'union entre les deux :

$$IoU = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- La **précision**, **rappel** et le **score F1** sont aussi des métriques utilisées.

Dans notre expérience avec SAM **nous allons utiliser IoU** pour comparer la ground truth mask et la prédiction et ainsi mesurer les performances et comparer avec un autre modèle.

4. Segment Anything Model

Avec cette introduction et informations de base sur la segmentation d'images, nous pouvons maintenant passer au modèle que nous avons étudié et à notre expérience.
Comme déjà mentionné, ce projet se base sur la compréhension et application du modèle que nous allons appeler SAM (de Segment Anything Model), introduit dans la prépublication [1].

L'article Segment Anything, et plus précisément, le projet Segment Anything a visé trois objectifs : une **tâche**, un **modèle** et un **dataset**. Essayons d'expliquer cela intuitivement.

Dans Segment Anything, l'équipe FAIR de Meta AI a introduit un **modèle de fondation** pour la segmentation d'image. Selon l'article en ligne [5], *un modèle de fondation est un modèle de grande taille, entraîné sur une grande quantité de données non étiquetées (généralement par apprentissage auto-supervisé). Le modèle résultant peut être adapté à un large éventail de* ***tâches en aval*** *("**downstream tasks**" en anglais).* Une tâche en aval est la tâche finale que l'on veut obtenir en entraînant un modèle. Par exemple, on peut entraîner un réseau de neurones pour reconnaître la race de chien sur un grand ensemble de photos de chiens, étiquetées avec les noms de races. La tâche en aval du modèle sera donc une classification et on pourra lui donner en entrée une photo et il donnera en résultat une des races des étiquettes apprises. Le but et la clé d'avoir un modèle de fondation est qu'il pourra servir à faire des généralisations : il ne sera pas destiné à une tâche en aval, mais il sera entraîné sur une grande quantité de données et à partir d'un ***prompt***[3], il pourra faire différentes tâches.

Tâche. Pour faire un tel modèle et type de tâches, les prompts pour utiliser SAM sont très variés : cela peut être des coordonnées sur l'image, équivalents à des clics (positifs ou négatifs, selon ce que l'on veut segmenter et le fond), cela peut être une *bounding box,* c'est à dire un rectangle qui borne l'objet à segmenter, ou bien un prompt de langage: une phrase. Voir la Figure A[4].

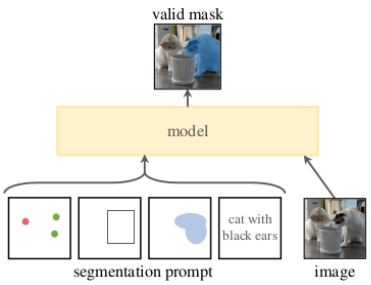


Figure A

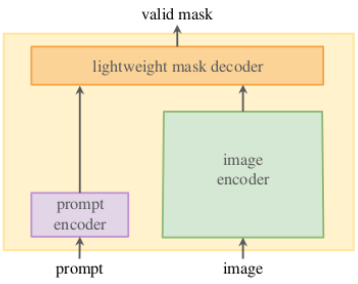


Figure B

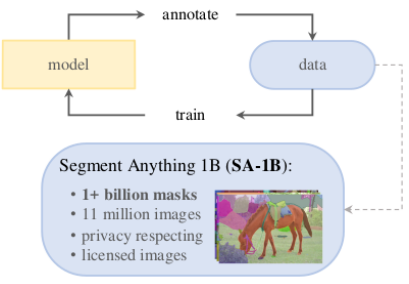
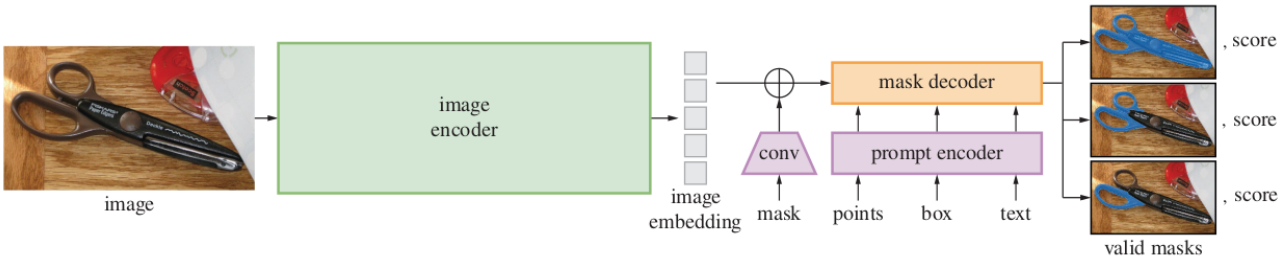


Figure C

[3] Invite de commande ou autre message destiné à l'utilisateur lui indiquant comment interagir avec un programme.

[4] Diagrammes pris de l'article [1]

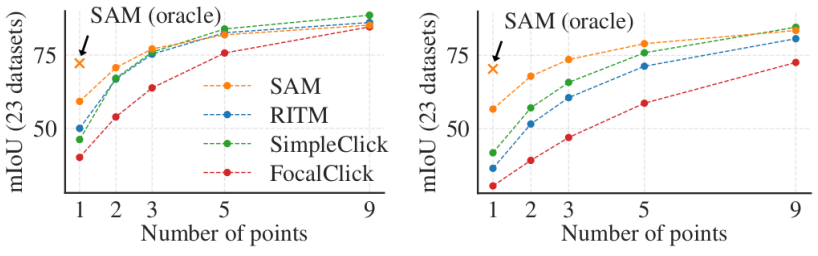
Modèle. Comme dit dans le paragraphe précédent, le modèle doit être flexible par rapport aux prompts. Il doit être rapide en plus et il doit être "conscient" de l'**ambiguïté**. En effet, les prompts données par un utilisateur ne vont pas toujours être clairs et objectifs. Par exemple, si je décide de sélectionner le T-shirt d'une personne et je clique sur son T-shirt, qui s'avère avoir un logo d'un chat dessus, ce n'est pas clair si je veux segmenter: la personne, son T-shirt ou le logo du chat sur son T-shirt. SAM gère donc l'ambiguïté en faisant multiples prédictions en sortie.



Sans entrer dans les détails des couches, ce diagramme et la Figure B illustrent l'architecture de SAM. En particulier les encodeurs et décodeurs sont construits au-dessus d'outils assez puissants et efficaces pour le traitement d'images et de textes comme un MAE (Masked Auto Encoder [6]), des Transformers ([7] par exemple).

Dataset. En utilisant leur modèle efficace dans une boucle de collecte de données, l'équipe FAIR de Meta AI a construit le plus grand ensemble de données de segmentation à ce jour. Avec plus d'**un milliard** de masks sur **11 millions d'images** sous licence et respectant la vie privée. Ce dataset est appelé SA-1B.
Dans cet étude, on s'aperçoit que les étiquettes, dans ce cas: les masks, ne sont pas en abondance sur internet. En effet, les données de pré-entrainemet pour les modèles de fondation en NLP (Natural Language Processing), se trouvent en abondance et d'accès facile sur internet. et si bien c'est naturel de trouver beaucoup d'images d'accès libre sur internet, les masks ne sont pas des données naturelles. La solution des auteurs de SAM est de créer ce qu'ils appellent un *data engine*, voir la Figure C. Les données initiales sont annotées à l'aide du modèle mais d'évaluation par des humains aussi, ensuite SAM est capable de générer automatiquement les masks pour des objets semblables et les annotateurs (humains) font le reste et finalement SAM peut générer automatiquement des masks sur presque tous les objets sur une seule image pour obtenir à peu près 100 masks par image.
Les chercheurs ont veillé à ce que l'origine de ces images soit équilibré géographiquement dans le monde et par rapport aux groupes de personnes qui apparaissent dans les photos.

Évaluation. SAM a été évalué dans des expériences avec 23 différents dataset de segmentation. Ce qui a été dans un sens un obstable pour le présent projet. En effet, il y a peu de données de segmentation (images + masks), donc cela n'a pas été facile de trouver un dataset que les auteurs n'aient pas utilisé pour notre expérience, on parlera de notre dataset plus en détail dans la section suivante (§4.2).
Les auteurs font des expériences avec ces 23 datasets de segmentation et avec plusieurs baselines. Un de leurs baselines, le principal est RITM [8] et c'est celui que l'on a choisi pour notre expérience.



(c) Center points (default) (d) Random points

Ces expériences ont montré que SAM a des meilleurs performances que RITM, en particulier, pour tous les datasets, en choisissant comme prompt soit un point au milieu, soit un point au hasard, voir

image ci dessus[5]. Les auteurs ont utilisé mIoU: la moyenne de l'indice de Jaccard, mentionné plus haut dans ce rapport, sur toutes les images. Ils ont également fait une étude humaine en laissant des personnes sélectionner les meilleurs masks prédites.

Plusieurs autres aspects et des expériences intéressantes peuvent être lues en détail dans l'article en question.

5. Expérience et évaluation avec notre dataset

Ayant abordé la segmentation d'image et le modèle SAM, nous sommes prêts à illustrer notre expérience. Il s'agît de tester les performances de SAM en l'appliquant, selon les consignes du projet, à un dataset qui n'ait pas été traité dans l'article original.
Nous avons:

- Trouvé un **dataset** pour mener notre expérience. C'est-à-dire, un ensemble d'images et leurs masks correspondants. Avec la condition que le dataset n'ait pas été utilisé dans l'article étudié.
- Choisi une **baseline** pour comparer avec le modèle étudié.
- Préparé les données et implémenté les deux modèles dans un notebook Jupyter. Nous avons finalement **prédit** des masks et comparé les **performances.**

Dans le reste de cette section nous décrivons cette démarche en détail.

5.1. Dataset

Comme déjà mentionné, les images annotées avec des masks ne sont pas abondantes sur internet, et les bons dataset disponibles ont été utilisés dans l'article de SAM pour évaluer le modèle. Malgré ces contraintes, nous avons trouvé un dataset adapté à notre objectif, qui est prédire des masks et les comparer.
Notre dataset ([9],[10]) est accessible sur la plateforme Kaggle. Malheureusement l'article d'origine n'est pas en libre accès donc nos connaissances sur la stratégie pour recueillir ces données ou générer les masks sont limitées.

Description des données. Ce dataset contient des photos en format `.png` de 9 différentes espèces de fruits de mer (poissons, crevettes). Les espèces sont les suivantes:

Espèce (nom original)	Espèce (nom en français)	Nombre d'images
Hourse Mackerel	Maquereau	50
Black Sea Sprat	*Clupeonella cultriventris*	50
Striped Red Mullet	Rouget de roche	50
Gilt-Head Bream	Dorade royale	50
Red Mullet	Rouget	50
Sea Bass	Bar commun	50
Shrimp	Crevette	50
Trout	Truite	50
Red Sea Bream	Dorade japonaise	50
		Total: 450

[5] Image prise de l'article [1].

Pour chacune des espèces il y a 1000 photos qui ont été augmentées (flip et rotation), mais en réalité viennent de 50 différentes photos pour chaque espèce (et les 50 masks de ground truth correspondantes). Une partie de notre travail a été de choisir ces 50 photos par espèce. En effet puisqu'on ne fera pas de *fine-tuning* nous n'avons pas besoin d'augmenter les données.

5.2. Baseline

La baseline choisie est le modèle que nous allons appeler RITM et qui a été introduit dans l'article [8]. La raison de l'avoir choisi est que c'est une des baselines dans l'article de SAM et que le code avec le modèle et l'implémentation est bien documenté et avec des exemples de code dans leur dépôt GitHub [11]. C'est un modèle de segmentation d'image interactif : le prompt pour segmentation est un clic, ou un ensemble de clics (coordonnées dans l'image).

5.3. Expérience et évaluation

Dans cette section nous expliquons en détail la démarche de notre expérience. Le code correspondant à cette expérience se trouve dans le notebook `segmentation_experience.ipynb` (lien) dans le dépôt GitHub du présent rapport [12].

a. Dans un premiers temps nous importons toutes les images de notre dataset dans un dictionnaire python avec clés chaque espèce de fruit de mer. Les valeurs contiennent les **images** et les **ground truth masks** comme des **tableaux numpy**, convertis grâce à la librairie cv2.

b. Ensuite nous importons des poids pour utiliser le modèle **RITM**, basés sur le code dans le notebook [13]. Pour leur évaluation, les auteurs de RITM ont écrit des scripts qui choisissent des points (des clics) automatiquement. Nous exécutons RITM pour prédire des masks sur toutes nos images, et nous gardons ces clics pour les utiliser pour la prédiction avec SAM.

   Dans la Figure 1 nous pouvons voir que RITM a choisi une suite de 2 clics au total. Cet algorithme continue à cliquer sur l'image jusqu'à obtenir un IoU qui est le but à obtenir. Une fois il dépasse ce but il s'arrête. Pour cette expérience nous avons choisi 0.9 comme but pour IoU et un maximum de 10 clics. La dernière prédiction sera le mask avec le meilleur score IoU (le plus proche de 1 atteint).

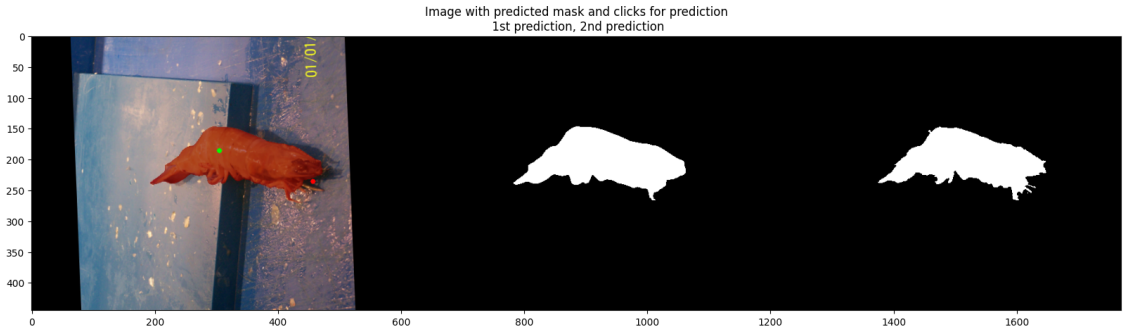


Figure 1: Prédiction avec RITM

c. Nous importons ensuite les poids du modèle **SAM** pré-entrainé et exécutons la prédiction sur nos images avec les clics donnés dans la partie b. par RITM.
   Rappelons que SAM est conscient de l'ambiguité d'un prompt (ici, un clic). La prédiction donne donc 3 masks ambiguës et prédit un score IoU estimé (que nous allons appeler **SIoU**, S comme SAM) pour chacune des maks prédites, voir la Figure 2.

Figure 2: Ambiguïté des prédictions avec SAM

Pour notre expérience, puisqu'il faut choisir un mask pour chacune des 450 images, nous choisissons celui avec le plus grand SIoU.

d. **Évaluation:** Maintenant nous sommes à disposition de:
   - Les images originales
   - Les masks de ground truth
   - Les masks prédites avec RITM dans la partie b.
   - Les masks prédites avec SAM dans la partie c.

Nous utilisons la métrique choisie: IoU entre les masks de ground truth et RITM et entre les masks de ground truth et SAM. Les tableaux suivants montrer quelques statistiques sur ce score. Rappelons que l'IoU mesure la ressemblence de deux ensembles, dans notre cas, des masks. Plus l'IoU est proche de la valeur 1, plus les ensembles ressemblent entre eux. **On vise donc un IoU proche de 1.**

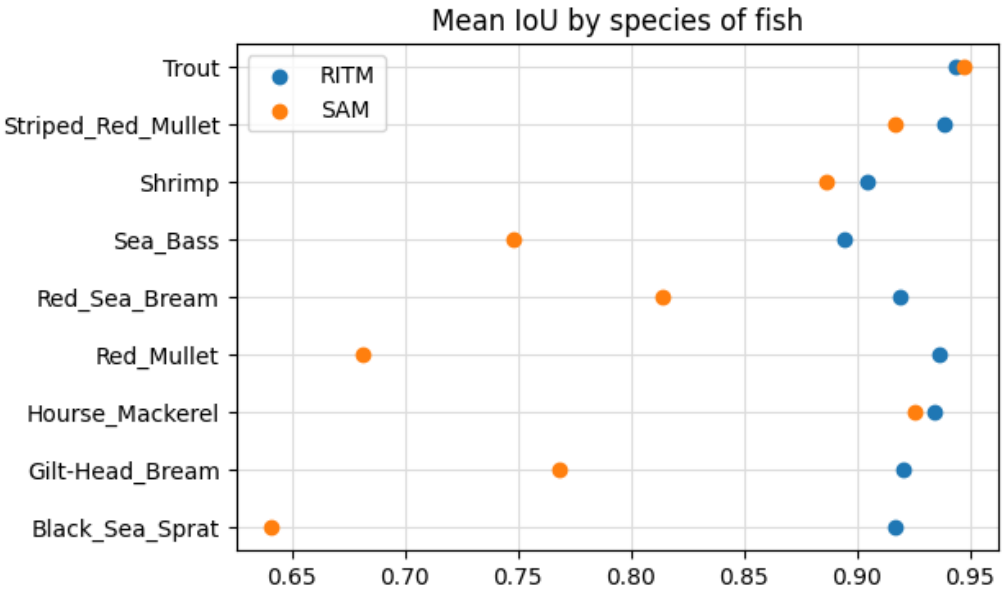
Quelques statistiques sur tout le dataset

	iou_RITM	iou_SAM
count	450.000000	450.000000
mean	0.922935	0.814156
std	0.043442	0.259960
min	0.526922	0.023829
25%	0.910158	0.823418
50%	0.925036	0.920395
75%	0.943318	0.958745
max	0.992594	0.987818

Moyenne de IoU par espèce

species	iou_RITM	iou_SAM
Black_Sea_Sprat	0.916618	0.640691
Gilt-Head_Bream	0.920119	0.768064
Hourse_Mackerel	0.934028	0.925069
Red_Mullet	0.936137	0.681345
Red_Sea_Bream	0.918904	0.813904
Sea_Bass	0.894340	0.748280
Shrimp	0.904728	0.886251
Striped_Red_Mullet	0.938162	0.916985
Trout	0.943377	0.946816

Moyenne de IoU par espèce: graphe



Le modèle SAM n'a pas les résultats attendus par rapport à RITM. Nous parlerons de nos hypothèses à propos dans les conclusions (§6).

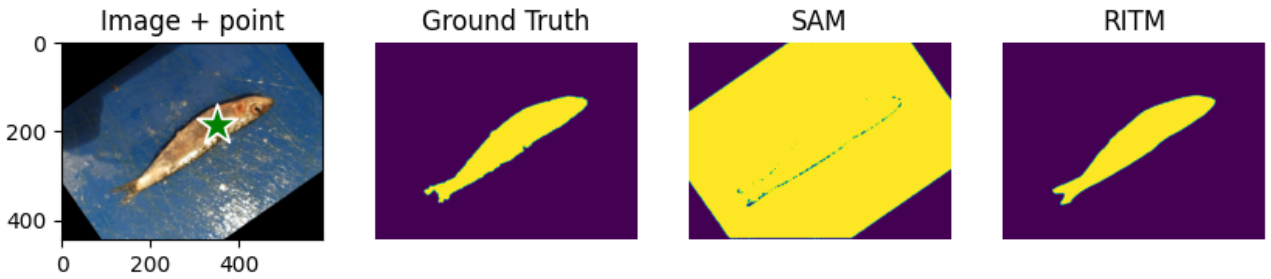
6. Conclusions

Contrairement aux attentes, dans cette expérience, nous obtenons de **meilleurs résultats** avec RITM (le modèle de base) qu'avec SAM (le modèle principal). De mon point de vue de non-expert, je pense que les raisons en sont les suivantes :

- Manque de qualité des étiquettes de l'ensemble de données.** Certains masques de vérité au sol sont de très mauvaise qualité par rapport à l'image originale. Nous ne savons pas comment ces masques ont été générés et malheureusement de la source de ce jeu de données n'est pas disponible gratuitement.
- Comme SAM est entraîné à effectuer des tâches ambiguës, nous n'avons choisi qu'un seul des masks prédits pour notre évaluation. Parfois, avec notre **unique** clic choisi, le masque choisi a sélectionné un "objet" correspondant à la quasi-totalité de l'image (voir les exemples ci-dessus).

Mais il est certain (voir l'exemple ci-dessus) que dans ce cas, le bon masque a également été prédit comme l'une des prédictions ambiguës. En effet, en regardant en détail quelques pas de notre boucle

de prédiction nous avons :



Dans ce cas, la mask prédit par SAM ne correpond pas à nos attentes. Si nous régardons de près les 3 prédictions ambiguës de SAM sur la même image, nous avons :

Effectivement il y a un de masks que aurait pu avoir un meilleur score. D'ailleurs quand on calcul ce sore (IoU) le résultat est `IoU = 0.96`.

La solution à ce problème pourrait être de choisir un ensemble de plusieurs clics au lieu d'un seul. Il s'agit d'une expérience facile qui peut être réalisée avec plus de temps.

Bibliographie

[1] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

[2] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(7), 3523-3542.

[3] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193-202.

[4] Chen, L. C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3640-3649).

[5] Modèle de fondation. (2023, février 4). *Wikipédia, l'encyclopédie libre*. Page consultée le 18:02, février 4, 2023 à partir de http://fr.wikipedia.org/w/index.php?title=Mod%C3%A8le_de_fondation&oldid=201085167.

[6] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000-16009).

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[8] Sofiiuk, K., Petrov, I. A., & Konushin, A. (2022, October). Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 3141-3145). IEEE.

[9] Ulucan, O., Karakaya, D., & Turkan, M. (2020, October). A large-scale dataset for fish segmentation and classification. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-5). IEEE.

[10] M. Turkan. (2020). A Large Scale Fish Dataset, Version 2. https://www.kaggle.com/datasets/crowww/a-large-scale-fish-dataset

[11] Dépôt GitHub de l'article [8] https://github.com/SamsungLabs/ritm_interactive_segmentation

[12] Dépôt GitHub du présent rapport https://github.com/ana-bernal/P7_preuve-concept

[13] Notebook d'exemple par les auteurs de l'article [8] https://github.com/SamsungLabs/ritm_interactive_segmentation/blob/master/notebooks/test_any_model.ipynb