Apprentissage de représentations d'images multimodales : texte / image via CLIP

Gabriel LUCCHINI

1.

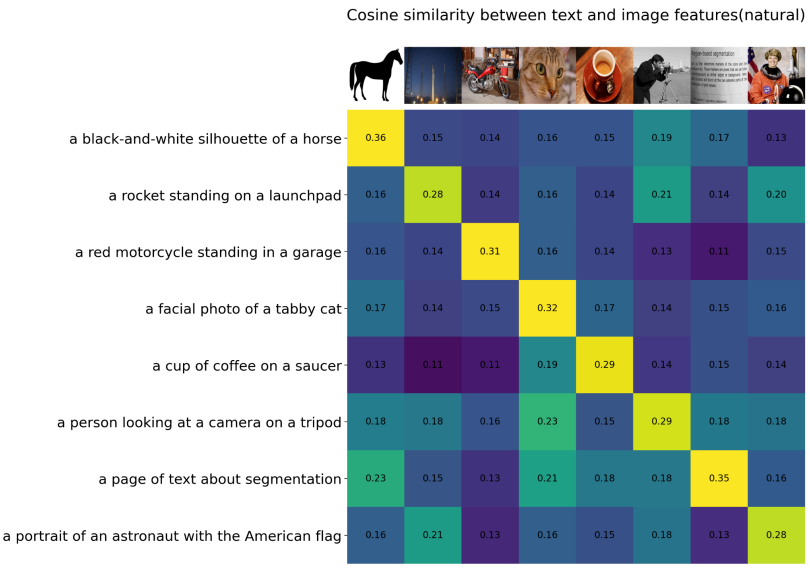
Voir code source.

2.

Les différents modèles de CLIP disponibles sont : RN50, RN101, RN50x4, RN50x16, RN50x64, ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14@336px.
On décide de choisir les modèles ViT-B/16, RN-50 et ViT-L/14 pour les évaluer et les confronter.
On calcule la similarité cosine pour des pairs d'images et textes (naturels).

Description + résultat de ViT-B/16 :

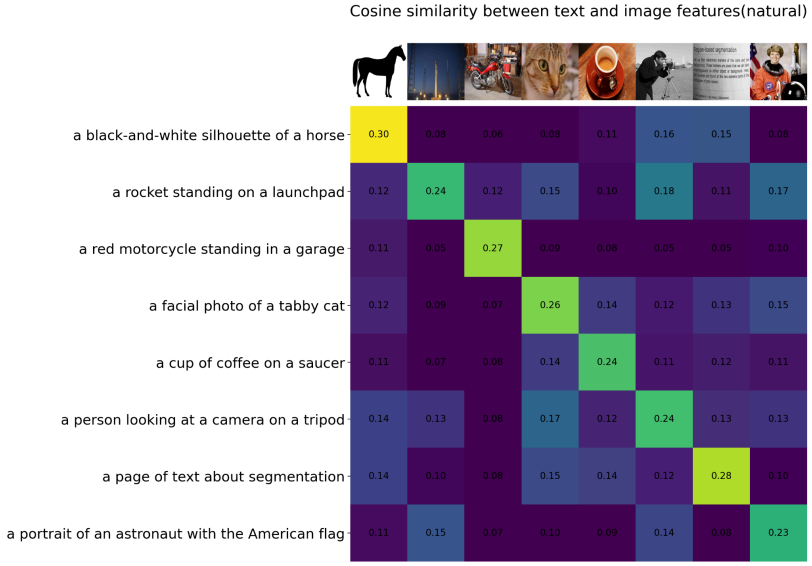
- Model parameters: 149,620,737
- Input resolution: 224
- Context length: 77
- Vocab size: 49408



Description + résultat de RN-50 :

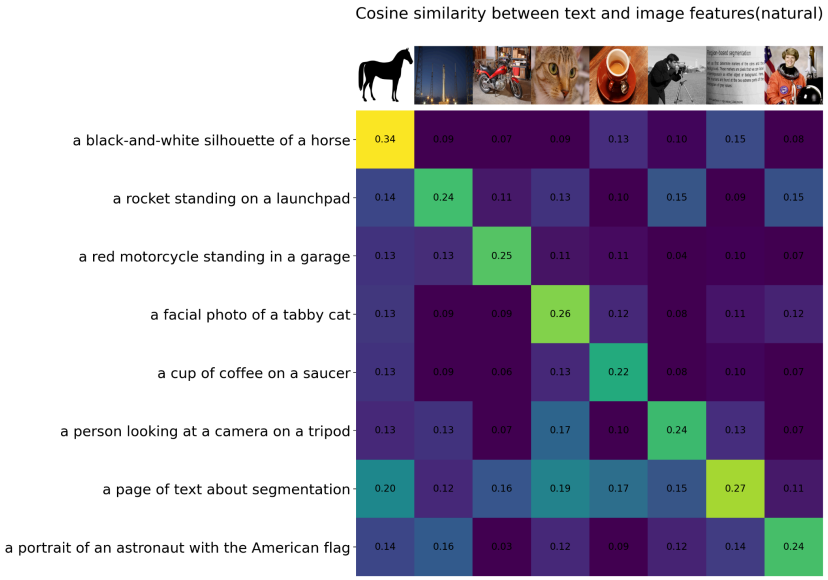
- Model_RN50 parameters: 102,007,137
- Input resolution: 224

- Context length: 77
- Vocab size: 49408



Description + résultat de ViT-L/14 :

- Model_vitl14 parameters: 427,616,513
- Input resolution: 224
- Context length: 77
- Vocab size: 49408

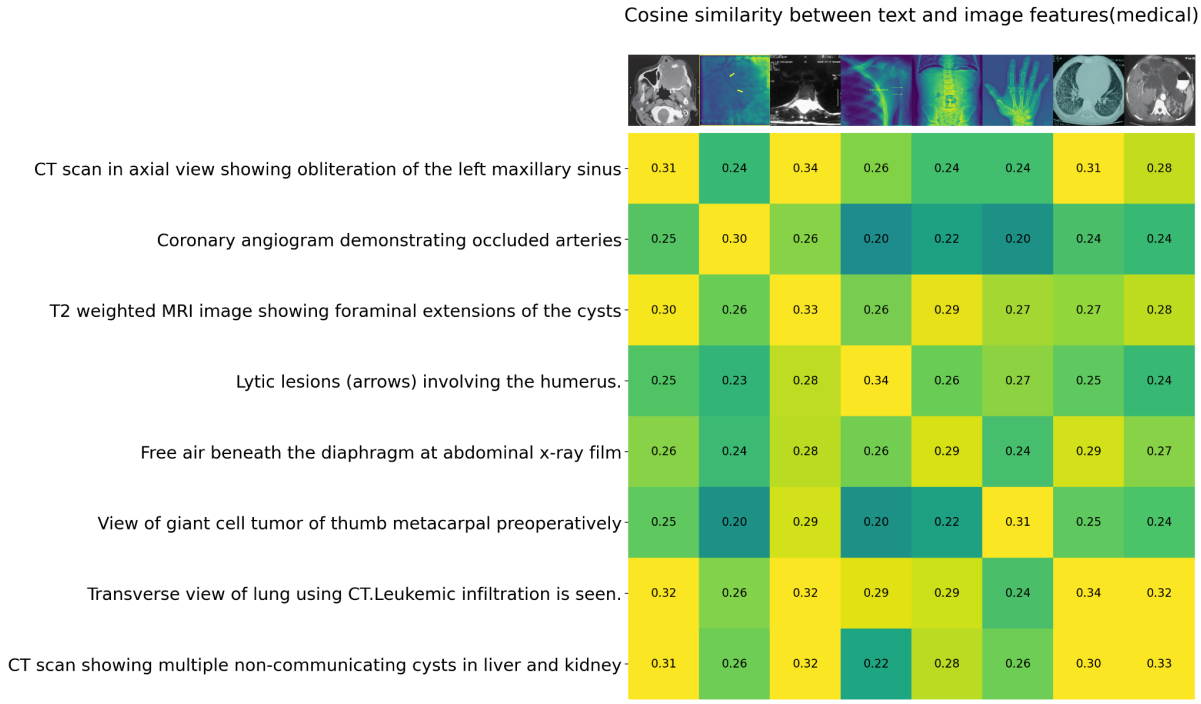


Résultats :

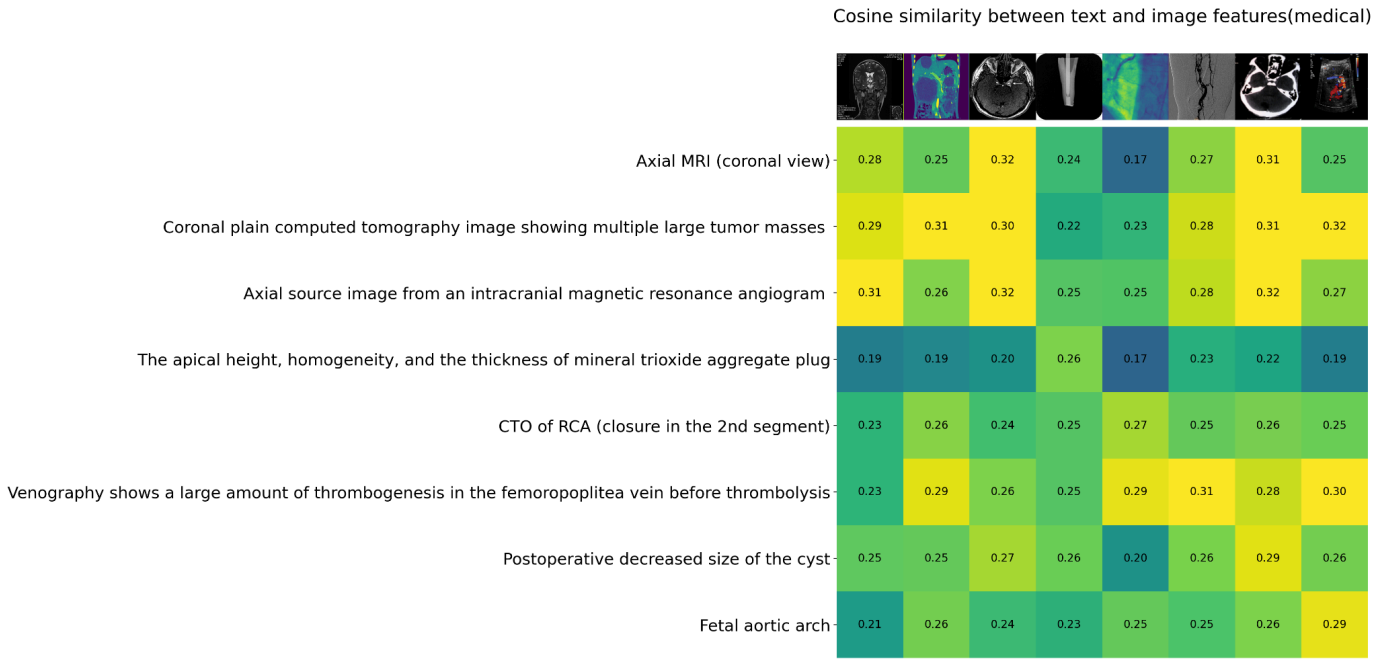
On observe que le modèle le plus performant pour notre tâche est le ViT-B/16. Les modèles RN-50 et ViT-L/14 ont des performances similaires. Malgré un grand nombre de paramètres, ViT-L/14 reste moins performant que ViT-B/16.

3.

Dans le code fourni, on avait déjà ces résultats de CLIP dans le domaine médical :



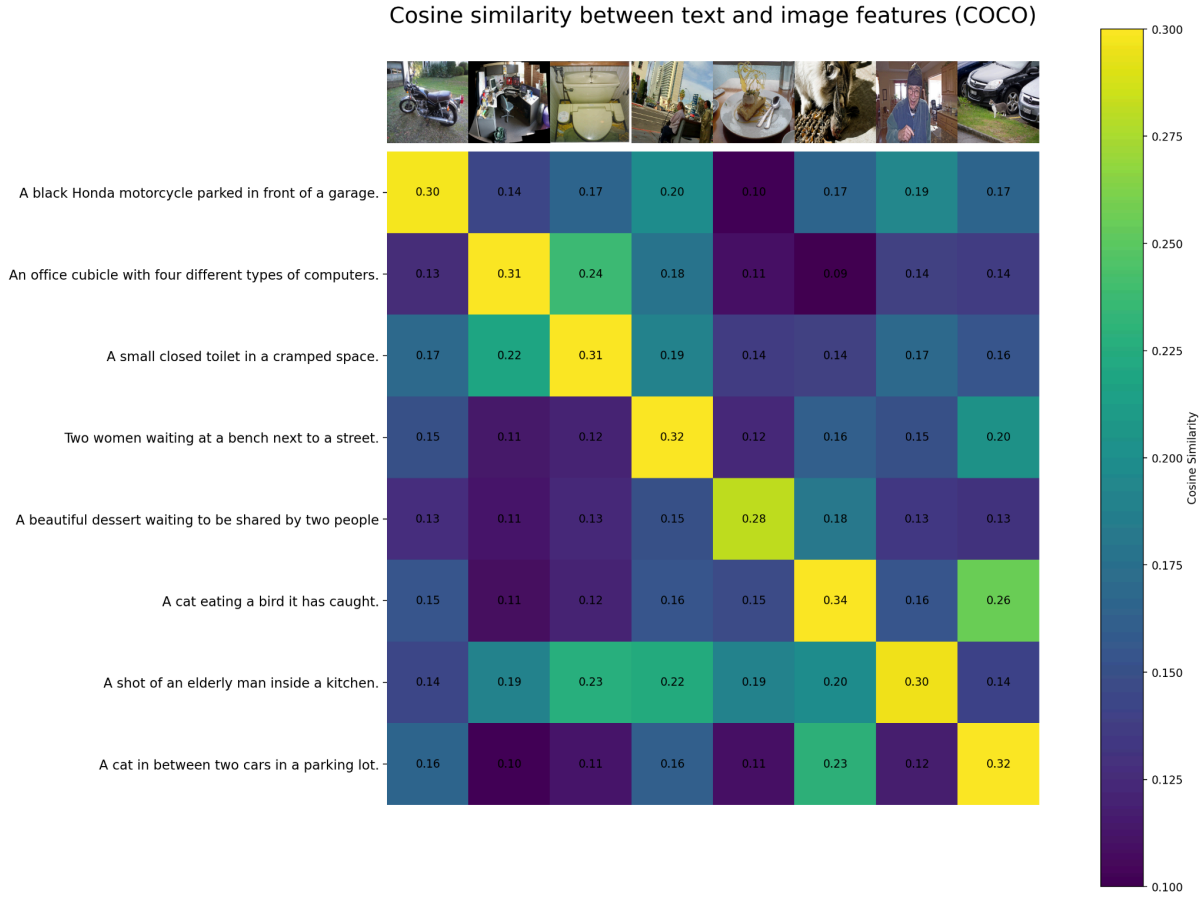
On choisit 8 autres images du jeu de données pour voir les résultats de CLIP :



On observe que CLIP n’est pas performant tant qu’il ne reçoit pas de fine-tuning en fonction du domaine.

4.

On décide de calculer les performances de CLIP sur le dataset COCO pour une tache de classification d'images naturelles :



On observe sur la matrice que COCO performe plutôt correctement sur des images naturelles.