

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

https://github.com/Leirbagg/TP_Multimodal
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.

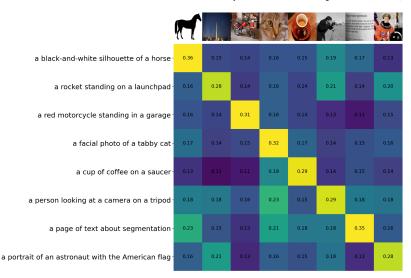
On calcule la similarité cosine pour des pairs d'images et textes (naturels).

Description + résultat de ViT-B/16 :

Model parameters: 149,620,737

Input resolution: 224Context length: 77Vocab size: 49408

Cosine similarity between text and image features(natural)



Description + résultat de RN-50 :

Model RN50 parameters: 102,007,137

- Input resolution: 224

Context length: 77Vocab size: 49408

Cosine similarity between text and image features(natural)

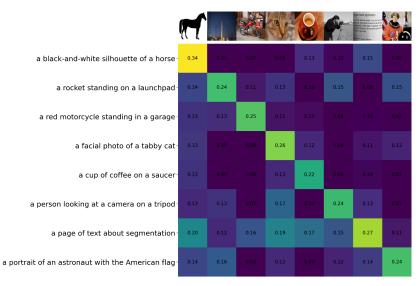
	M			29%	6	6	Region-based segmentation at a fee deepen private it to also act for more than these segments or all for more than the privation of the feet of the company of	
a black-and-white silhouette of a horse	0.30	0.08				0.16	0.15	0.08
a rocket standing on a launchpad		0.24	0.12	0.15		0.18		0.17
a red motorcycle standing in a garage			0.27					0.10
a facial photo of a tabby cat-			0.07	0.26	0.14		0.13	0.15
a cup of coffee on a saucer			0.08	0.14	0.24	0.11		0.11
a person looking at a camera on a tripod	0.14	0.13		0.17	0.12	0.24	0.13	0.13
a page of text about segmentation	0.14			0.15	0.14	0.12	0.28	0.10
a portrait of an astronaut with the American flag	0.11	0.15	0.07	0.10	0.09	0.14	0.08	0.23

Description + résultat de ViT-L/14 :

- Model_vitl14 parameters: 427,616,513

Input resolution: 224Context length: 77Vocab size: 49408

Cosine similarity between text and image features(natural)

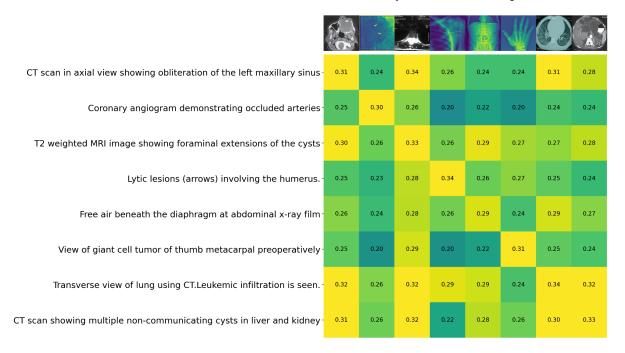


Résultats:

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.

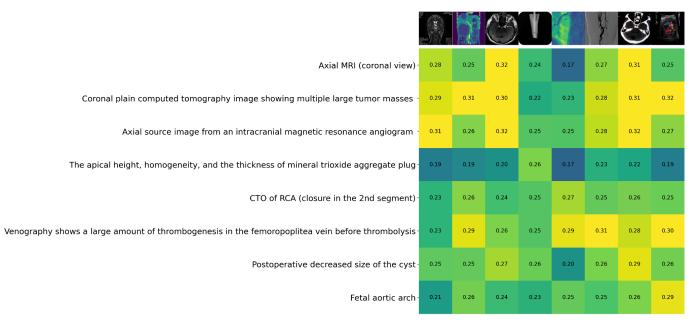
3. Dans le code fourni, on avait déjà ces résultats de CLIP dans le domaine médical :

Cosine similarity between text and image features(medical)



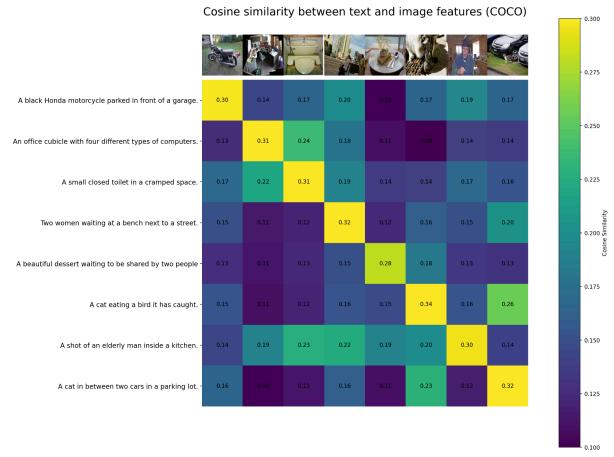
On choisit 8 autres images du jeu de données pour voir les résultats de CLIP :

Cosine similarity between text and image features(medical)



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.