Rapport TP1 Multimodalité

Hamza Omari[1*]

[1*]UFR Mathématiques et Informatique, Université Paris-Cité, Paris, 75006, France.

Corresponding author(s). E-mail(s): hamza.omari@etu.u-paris.fr;

1- familiarisation avec le notebook

Le notebook montre l'utilisation des modèles CLIP de la bibliothèque clip. Deux familles de modèles sont présentes : ResNet et ViT. Un jeu de données composé de paires (image, description) est utilisé. En calculant la similarité, on peut évaluer les performances des espaces d'embeddings et réaliser un fine-tuning sur le dataset ROCO. Le notebook présente ensuite l'utilisation du modèle après l'entraînement, son évaluation, ainsi qu'un exemple d'évaluation sur une tâche de classification.

Evaluations des modèles clips

ils existent deux familles de modèles présentès: ResNet et ViT RN50, RN101, RN50x4, RN50x16, RN50x64
ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14@336px

Pour une évaluation quantitative, j'ai d'abord opté pour une méthode simple : calculer les similarités image–texte, les visualiser sous forme de matrice, puis analyser la **trace** de cette matrice. La trace indique directement à quel point le modèle parvient à maximiser la similarité entre chaque image et son texte associé.

Cependant, en visualisant les matrices de similarités, j'ai constaté que certains modèles affichent une **bonne valeur de trace**, mais restent globalement **mauvais** : ils n'arrivent pas à séparer correctement les images et les textes non associés, ce qui se traduit par des valeurs off–diagonales élevées. Comme dans le cas du modèle RN101, voir la figure 1 Ainsi, la trace seule est insuffisante pour mesurer la qualité globale du modèle.

Pour cela, l'évaluation que j'ai retenue consiste à prendre en compte **à la fois** la diagonale (paires correctes) et les éléments hors–diagonale (paires incorrectes), selon la formule suivante :

$$\text{score} = \underbrace{\frac{\text{trace}(S)}{N}}_{\text{moyenne des similarités correctes}} - \underbrace{\frac{\sum S - \text{trace}(S)}{N^2 - N}}_{\text{moyenne des similarités incorrectes}} . \tag{1}$$

où S désigne la matrice de similarité et N le nombre d'images (et donc de textes associés).

Ce score n'augmente que lorsque :

- les similarités correctes (diagonale) augmentent,
- et les similarités incorrectes (hors–diagonale) diminuent.

Il constitue donc une mesure beaucoup plus fiable et robuste que la trace seule pour évaluer la performance globale d'un modèle CLIP.

les matrices de similarités calculées, pour les autres différents modèles se retrouve en bas de ce document.

Nous classons ensuite les modèles par ordre décroissant de performance, selon la métrique définie par l'équation 1. Les résultats obtenus sont les suivants

Fig. 1: Matrice de similarité pour le modèle RN101

Model	Score
ViT-L/14@336px	0.094371
ViT-L/14	0.092693
RN50x64	0.091421
RN50x16	0.081308
RN50x4	0.061293
ViT-B/16	0.056187
ViT-B/32	0.054636
RN101	0.054573
RN50	0.052011

Fig. 2: Evaluations

Performance sur d'aures images médicales

Pour cette tâche, nous sélectionnons des images différents et nous utilisons ChatGPT pour les annoter (c'est-à-dire les décrire). Ensuite, de la même manière que précédemment, nous calculons la matrice de similarité. voir figure 3

Évaluation sur une tâche de classification

Pour cette partie, nous utilisons un dataset de Hugging Face, *Food101*. Nous sélectionnons quelques images et définissons des descriptions textuelles correspondant aux différentes classes. Ensuite, nous calculons la similarité entre chaque image et l'ensemble des descriptions des classes choisies. Les logits obtenus sont ensuite

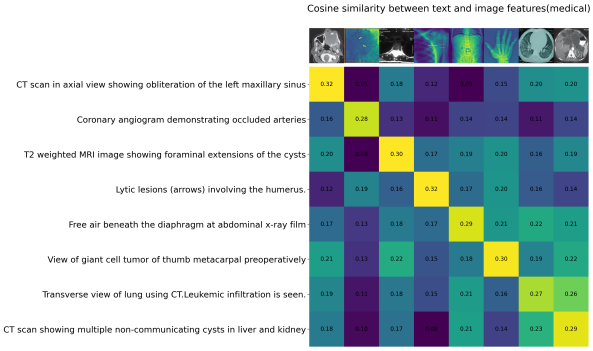


Fig. 3: Evaluations sur d'autres images

passés dans une fonction *softmax*, ce qui nous permet de visualiser la distribution des probabilités et d'interpréter les prédictions du modèle.

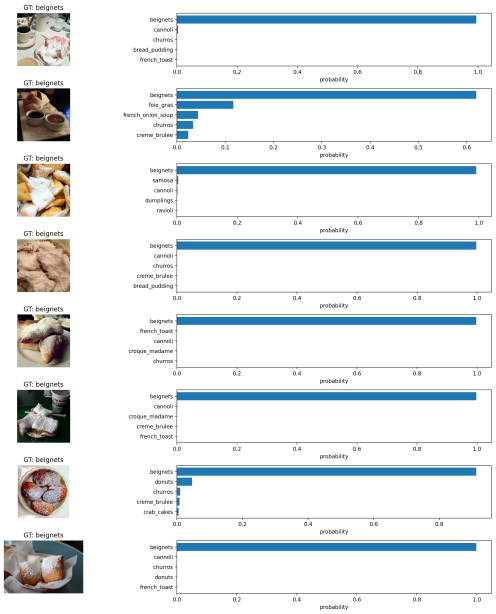


Fig. 4: Evaluations sur la classif

On voit bien que le modèle arrive a bien associer l'image à ça vraie classe.

Code source

: Continuation sur le même notebook. GitHub

Cosine similarity for RN50

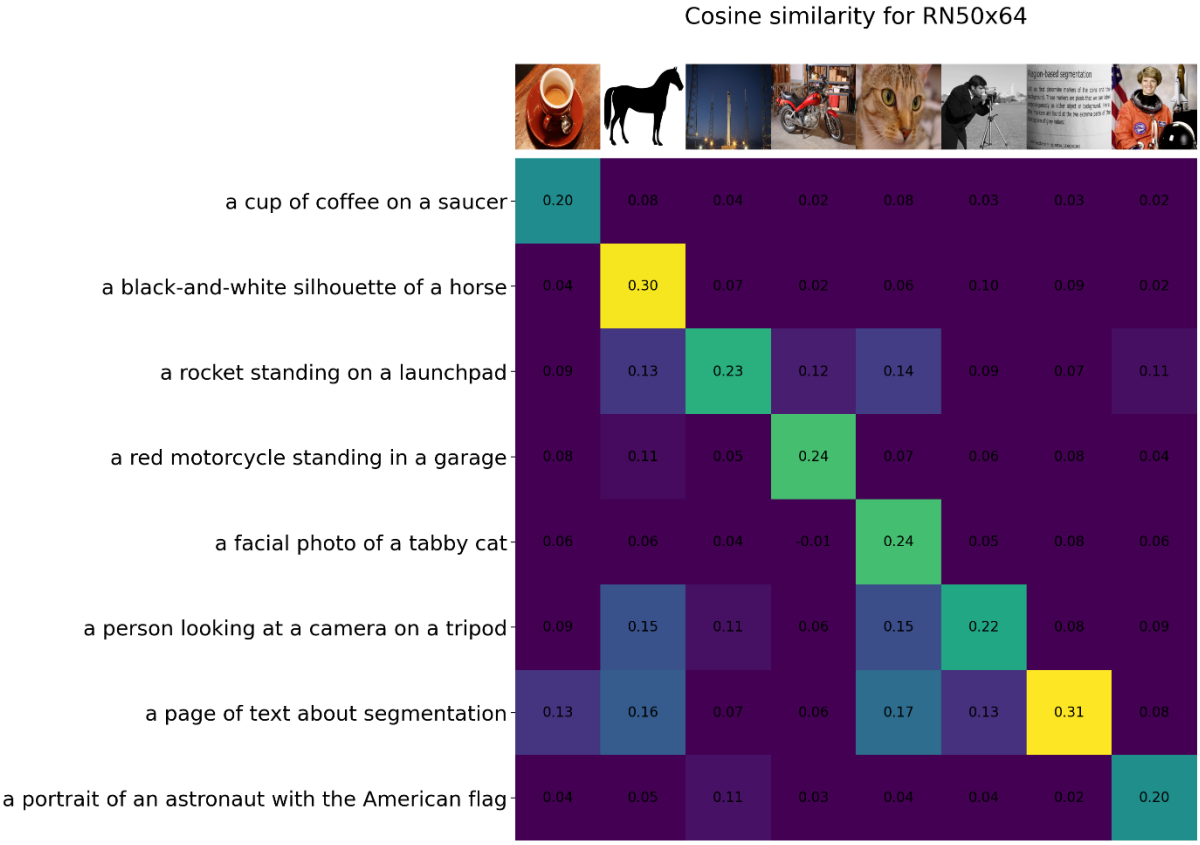
a cup of coffee on a saucer	0.24	0.11	0.07	0.08	0.14	0.11	0.12	0.11
a black-and-white silhouette of a horse	0.11	0.30	0.08	0.06	0.08	0.16	0.15	0.08
a rocket standing on a launchpad	0.10	0.12	0.24	0.12	0.15	0.18	0.11	0.17
a red motorcycle standing in a garage	0.08	0.11	0.05	0.27	0.09	0.05	0.05	0.10
a facial photo of a tabby cat	0.14	0.12	0.09	0.07	0.26	0.12	0.13	0.15
a person looking at a camera on a tripod	0.12	0.14	0.13	0.08	0.17	0.24	0.13	0.13
a page of text about segmentation	0.14	0.14	0.10	0.08	0.15	0.12	0.28	0.10
a portrait of an astronaut with the American flag	0.09	0.11	0.15	0.07	0.10	0.14	0.08	0.23

Cosine similarity for RN50x4

a cup of coffee on a saucer	0.42	0.27	0.23	0.22	0.30	0.22	0.24	0.22
a black-and-white silhouette of a horse	0.26	0.45	0.23	0.25	0.26	0.30	0.29	0.24
a rocket standing on a launchpad	0.28	0.29	0.40	0.27	0.31	0.30	0.28	0.30
a red motorcycle standing in a garage	0.26	0.25	0.22	0.40	0.28	0.22	0.24	0.24
a facial photo of a tabby cat	0.27	0.26	0.21	0.21	0.41	0.23	0.25	0.20
a person looking at a camera on a tripod	0.29	0.31	0.30	0.25	0.34	0.40	0.31	0.30
a page of text about segmentation	0.31	0.30	0.23	0.24	0.33	0.28	0.47	0.25
a portrait of an astronaut with the American flag	0.23	0.23	0.29	0.23	0.29	0.26	0.26	0.37

Cosine similarity for RN50x16

a cup of coffee on a saucer	0.30	0.17	0.12	0.16	0.18	0.16	0.14	0.14
a black-and-white silhouette of a horse	0.14	0.36	0.12	0.15	0.13	0.15	0.20	0.10
a rocket standing on a launchpad	0.16	0.20	0.28	0.20	0.18	0.19	0.17	0.21
a red motorcycle standing in a garage	0.16	0.18	0.13	0.33	0.16	0.11	0.14	0.13
a facial photo of a tabby cat	0.16	0.18	0.14	0.15	0.29	0.12	0.13	0.12
a person looking at a camera on a tripod	0.18	0.20	0.19	0.15	0.24	0.29	0.21	0.17
a page of text about segmentation	0.20	0.24	0.14	0.14	0.24	0.17	0.37	0.14
a portrait of an astronaut with the American flag	0.10	0.11	0.19	0.11	0.14	0.15	0.16	0.27


Cosine similarity for RN50x64
a cup of coffee on a saucer
a black-and-white silhouette of a horse
a rocket standing on a launchpad
a red motorcycle standing in a garage
a facial photo of a tabby cat
a person looking at a camera on a tripod
a page of text about segmentation
a portrait of an astronaut with the American flag
0.20 0.08 0.04 0.02 0.08 0.03 0.03 0.02
0.04 0.30 0.07 0.02 0.06 0.10 0.09 0.02
0.09 0.13 0.23 0.12 0.14 0.09 0.07 0.11
0.08 0.11 0.05 0.24 0.07 0.06 0.08 0.04
0.06 0.06 0.04 -0.01 0.24 0.05 0.08 0.06
0.09 0.15 0.11 0.06 0.15 0.22 0.08 0.09
0.13 0.16 0.07 0.06 0.17 0.13 0.31 0.08
0.04 0.05 0.11 0.03 0.04 0.04 0.02 0.20

Cosine similarity for ViT-B/16

a cup of coffee on a saucer	0.29	0.13	0.11	0.11	0.19	0.14	0.15	0.14
a black-and-white silhouette of a horse	0.15	0.36	0.15	0.14	0.16	0.19	0.17	0.13
a rocket standing on a launchpad	0.14	0.16	0.28	0.14	0.16	0.21	0.14	0.20
a red motorcycle standing in a garage	0.14	0.16	0.14	0.31	0.16	0.13	0.11	0.15
a facial photo of a tabby cat	0.17	0.17	0.14	0.15	0.32	0.14	0.15	0.16
a person looking at a camera on a tripod	0.15	0.18	0.18	0.16	0.23	0.29	0.18	0.18
a page of text about segmentation	0.18	0.23	0.15	0.13	0.21	0.18	0.35	0.16
a portrait of an astronaut with the American flag	0.15	0.16	0.21	0.13	0.16	0.18	0.13	0.28

Cosine similarity for ViT-B/32

a cup of coffee on a saucer	0.29	0.15	0.12	0.12	0.18	0.17	0.14	0.15
a black-and-white silhouette of a horse	0.15	0.35	0.15	0.17	0.15	0.21	0.17	0.11
a rocket standing on a launchpad	0.14	0.17	0.30	0.16	0.18	0.20	0.17	0.19
a red motorcycle standing in a garage	0.13	0.16	0.16	0.32	0.15	0.12	0.14	0.15
a facial photo of a tabby cat	0.17	0.15	0.12	0.12	0.31	0.16	0.12	0.12
a person looking at a camera on a tripod	0.14	0.20	0.21	0.16	0.21	0.30	0.19	0.19
a page of text about segmentation	0.20	0.20	0.16	0.16	0.20	0.20	0.35	0.15
a portrait of an astronaut with the American flag	0.15	0.16	0.22	0.15	0.17	0.17	0.13	0.28

Cosine similarity for ViT-L/14

a cup of coffee on a saucer	0.22	0.13	0.09	0.06	0.13	0.08	0.10	0.07
a black-and-white silhouette of a horse	0.13	0.34	0.09	0.07	0.09	0.10	0.15	0.08
a rocket standing on a launchpad	0.10	0.14	0.24	0.11	0.13	0.15	0.09	0.15
a red motorcycle standing in a garage	0.11	0.13	0.13	0.25	0.11	0.04	0.10	0.07
a facial photo of a tabby cat	0.12	0.13	0.09	0.09	0.26	0.08	0.11	0.12
a person looking at a camera on a tripod	0.10	0.13	0.13	0.07	0.17	0.24	0.13	0.07
a page of text about segmentation	0.17	0.20	0.12	0.16	0.19	0.15	0.27	0.11
a portrait of an astronaut with the American flag	0.09	0.14	0.16	0.03	0.12	0.12	0.14	0.24

Cosine similarity for ViT-L/14@336px

a cup of coffee on a saucer	0.24	0.13	0.10	0.06	0.13	0.08	0.11	0.08
a black-and-white silhouette of a horse	0.13	0.35	0.10	0.07	0.10	0.11	0.15	0.07
a rocket standing on a launchpad	0.10	0.14	0.25	0.12	0.13	0.17	0.11	0.16
a red motorcycle standing in a garage	0.11	0.14	0.12	0.27	0.11	0.05	0.09	0.07
a facial photo of a tabby cat	0.12	0.13	0.10	0.09	0.26	0.08	0.11	0.12
a person looking at a camera on a tripod	0.10	0.12	0.15	0.08	0.17	0.23	0.15	0.08
a page of text about segmentation	0.16	0.19	0.12	0.15	0.19	0.15	0.30	0.11
a portrait of an astronaut with the American flag	0.10	0.15	0.17	0.05	0.12	0.13	0.13	0.25