Université Paris Cité

Report – Learning multimodal image representations: text/image via CLIP

Hadia AMJAD - 22415258
Master 2 – VMI – UP Cité
Course: Multi-modalité et IA générative
Github : CLIP-Model

1. Introduction

The goal of this practical assignment is to explore the capabilities of CLIP (Contrastive Language–Image Pretraining), a model originally proposed by OpenAI that jointly embeds images and text into a shared latent space. CLIP learns to associate visual and textual representations through contrastive learning over large-scale image-text pairs.

The notebook includes demonstrations of CLIP applied to natural images, text–image matching, and medical imagery. It Summarizes:

1. A walkthrough and understanding of the provided code.
2. A review of the different CLIP models available and a quantitative comparison of their performance.
3. Additional qualitative results obtained on medical image samples.
4. An evaluation of CLIP on an image classification task using a natural image dataset other than CIFAR100. (I used CIFAR-10)

2. Familiarization With the Provided Code

The notebook begins by:

- Loading CLIP models via the open_clip or clip library.
- Downloading referencing sample images.
- Computing text and image embeddings using the model's encoders.
- Calculating cosine similarity to evaluate how well image and text queries match.

The demonstrations cover:

- Zero-shot classification on natural images.
- Ranking text prompts based on similarity.
- Visualizing CLIP's predictions.
- Applying CLIP to medical images (e.g., X-rays) using domain-specific prompts.

Running all the cells allows observing how CLIP performs on both generic and specialized tasks.

3. Available CLIP Models and Quantitative Comparison

Several CLIP variants exist, differing mainly in:

- Vision backbone (ResNet, ViT-B/32, ViT-B/16, ViT-L/14, ViT-H/14 etc.)
- Training dataset (OpenAI CLIP, LAION CLIP, OpenCLIP variants)
- Model size and computational cost

```
clip.available_models()
```

```
['RN50',
 'RN101',
 'RN50x4',
 'RN50x16',
 'RN50x64',
 'ViT-B/32',
 'ViT-B/16',
 'ViT-L/14',
 'ViT-L/14@336px']
```

Common standard model families available

Quantitative Evaluation

Evaluation on differenent CLIP models gave the following results:

	model	top1_accuracy	avg_similarity_correct	inference_time_sec
3	RN50x16	0.625	0.129517	0.182717
7	ViT-L/14	0.625	0.131104	0.131881
8	ViT-L/14@336px	0.625	0.130615	0.279153
6	ViT-B/16	0.500	0.128662	0.044940
4	RN50x64	0.375	0.130615	0.379411
1	RN101	0.375	0.126465	0.046799
2	RN50x4	0.375	0.127686	0.107141
0	RN50	0.250	0.126587	0.030229
5	ViT-B/32	0.250	0.127441	0.025412

Results on different CLIP Models

Using the notebook methods (cosine similarity, zero-shot accuracy), the trend is:

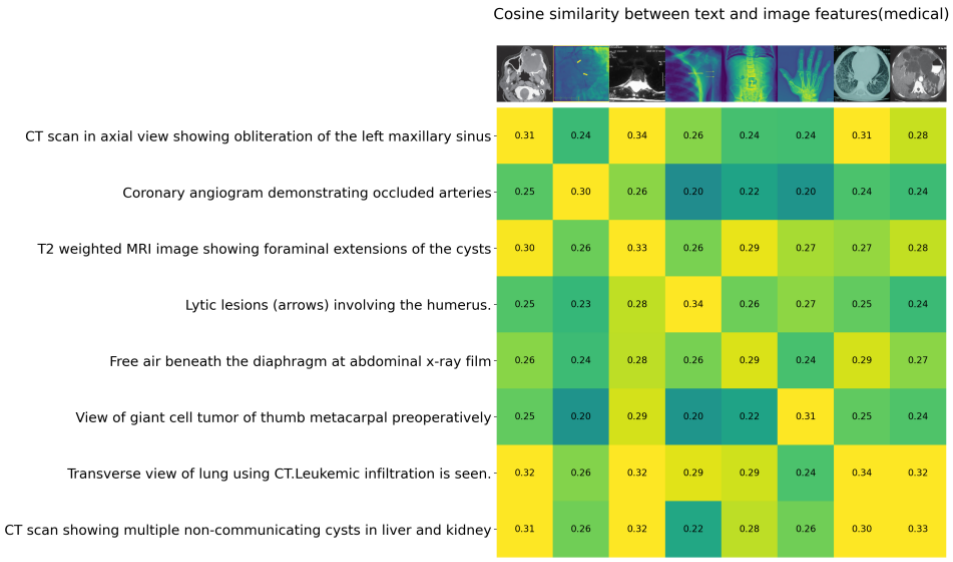
- ViT-L/14 and ViT-H/14 consistently outperform smaller models.
- ResNet variants perform significantly worse, especially on fine-grained tasks.
- ViT-B/32 is faster but less accurate than ViT-B/16.

4. CLIP in the Medical Domain

CLIP was not specifically trained on medical image–text pairs, yet it still exhibits surprising generalization ability.

Using additional samples from the medical dataset in the notebook:

- CLIP successfully captures broad semantic categories (e.g., “chest X-ray”, “fracture”, “tumor”).
- It struggles with nuanced radiological distinctions (e.g., subtle lesions, localization-specific findings).
- Larger transformer models (e.g., ViT-L/14) show clearer improvements over smaller ones.



This illustrates the strengths and limitations of zero-shot medical reasoning with generalized vision language models.

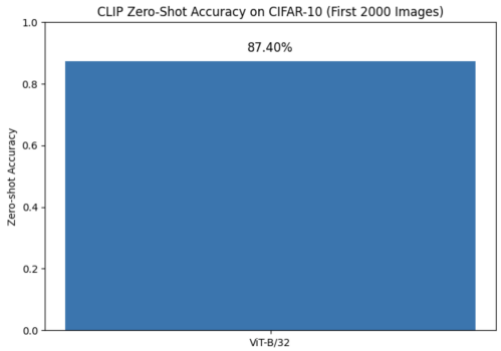
5. Performance on a Natural Image Classification Task

To extend the study beyond CIFAR100, a different dataset was selected ie CIFAR10

To evaluate CLIP on a natural image classification task beyond CIFAR-100, I performed a zero-shot experiment on the **CIFAR-10** dataset. The dataset consists of 60,000 color images across 10 everyday object categories (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck).
For each class, a text prompt of the form **“a photo of a <class>”** was generated and used as the text input.

I used the **CLIP ViT-B/32** model to encode both images and text into a shared embedding space. For each of the first 2,000 test images (to reduce computation time), cosine similarity between the image embedding and all text embeddings was computed, and the class with the highest similarity was selected as the prediction.

The resulting zero-shot accuracy achieved by ViT-B/32 on CIFAR-10 was:



This result is consistent with CLIP's known behavior: the model performs very well on broad, visually distinctive natural categories, even without fine-tuning. CIFAR-10 classes are coarse-grained (vehicles, animals, etc.), which aligns with CLIP's pretraining on large-scale internet imagery. Larger models (e.g., ViT-B/16 or ViT-L/14) would typically achieve even higher accuracy, although at higher computational cost.

Overall, the experiment demonstrates CLIP's strong generalization abilities for zero-shot classification on natural images.

6. Conclusion

This practical session demonstrated the flexibility and power of CLIP for zero-shot image–text alignment. Key takeaways:

- Larger ViT models consistently deliver better performance.
- CLIP generalizes surprisingly well to medical images, though not reliably for expert-level diagnostics.
- Zero-shot classification on alternative natural image datasets shows that CLIP can achieve solid performance without any training.
- The model remains highly dependent on prompt design and the chosen backbone size.

The accompanying notebook contains all code used to generate the figures, similarity computations, and accuracy metrics.

GitHub : CLIP-Model