Reproduction d'un dataset d'images

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

2026

Dans le domaine de la vision par ordinateur, CLIP [1] (Radford et al. 2021) est l'un des modèles de référence. Grâce à son entraînement multimodal texte-image, il permet en apprentissage zero-shot d'atteindre des performances état-de-lart sur des datasets traditionnellement traités par de l'apprentissage fermé. Le dataset utilisé dans l'article original et pour le modèle publiquement disponible est malheureusement fermé, on connaît simplement sa taille (400 millions de paires texte-image) et une description sommaire du processus d'acquisition.

Depuis des travaux on été entrepris (Cherti et al. 2023) pour reconstruire un dataset similaire permettant de reproduire l'entraînement de CLIP et de publier le modèle OpenCLIP [2] basé sur ce dataset ouvert. Dans ce cadre, deux datasets ont été publiés, de taille 400 millions [3] et 5 milliards [4] (Schuhmann et al. 2021, 2022).

Les objectifs du projet sont les suivants :

- explorer la littérature concernant les modèles vision-langage (VLM) ;
- reconstruire un dataset texte-image en utilisant la méthologie proposée pour LAION400m ;
- reproduire les expériences de l'article ;
- étudier l'impact de la taille et de la qualité des datasets sur les performances des modèles.

Bibliographie

Cherti, Mehdi, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. "Reproducible Scaling Laws for Contrastive Language-Image Learning." In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–29. https://doi.org/10.1109/CVPR52729.2023.00276.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." February 26, 2021. https://doi.org/10.48550/arXiv.2103.00020.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, et al. 2022. "LAION-5B : An Open Large-Scale Dataset for Training Next Generation Image-Text Models." October 16, 2022. https://doi.org/10.48550/arXiv.2210.08402.

Schuhmann, Christoph, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. "LAION-400M : Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs." November 3, 2021. https://doi.org/10.48550/arXiv.2111.02114.

1. https://openai.com/index/clip/
2. https://github.com/mlfoundations/open_clip
3. https://laion.ai/blog/laion-400-open-dataset/
4. https://laion.ai/blog/laion-5b/