1. Dataset

Link: [From Pixels to Prose: A Large Dataset of Dense Image Captions (arxiv.org)](https://arxiv.org/html/2406.10328v1#:~:text=PixelProse%20is%20a%20dataset%20that%20addresses)

Link download: [tomg-group-umd/pixelprose at main (huggingface.co)](https://huggingface.co/datasets/tomg-group-umd/pixelprose/tree/main/data)

PixelProse là một tập dữ liệu giải quyết các điểm yếu của bộ dữ liệu văn bản thay thế hiện có cho các ứng dụng ngôn ngữ thị giác và được thiết kế để sử dụng làm tài sản độc lập hoặc kết hợp với tái cấu trúc LLM. Nó chứa các chú thích chi tiết dài, chi tiết và bao gồm một loạt các thuộc tính hình ảnh quan trọng đối với Mô hình ngôn ngữ tầm nhìn (VLM) và đào tạo mô hình khuếch tán.

Tổng số cặp ảnh chú thích: 16,9 triệu

* 6,5 triệu cặp trong CommonPool chứa một nhóm lớn các cặp hình ảnh-văn bản từ CommonCrawl, được phân phối dưới dạng danh sách các cặp url-văn bản theo Giấy phép CC-BY-4.0
* 9,1 triệu cặp trong CC12M bao gồm hình ảnh được thu thập trên web và các cặp văn bản thay thế. Tập dữ liệu được quản lý bằng cả bộ lọc dựa trên hình ảnh và văn bản.
* 1,2 triệu cặp trong RedCaps được quản lý từ Reddit. Nó cặp hình ảnh-văn bản từ 350 subreddit khác nhau, được lọc để chọn ảnh chung và giảm thiểu số lượng người

1. Model cho việc train

VinVL:

* [VinVL/docs at main · pzzhang/VinVL (github.com)](https://github.com/pzzhang/VinVL/tree/main/docs)
* [VinVL: Revisiting Visual Representations in Vision-Language Models | Papers With Code](https://paperswithcode.com/paper/vinvl-making-visual-representations-matter-in)
* [VinVL: Revisiting Visual Representations in Vision-Language Models (thecvf.com)](https://openaccess.thecvf.com/content/CVPR2021/papers/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.pdf)

DenseCap:

* [jcjohnson/densecap: Dense image captioning in Torch (github.com)](https://github.com/jcjohnson/densecap)
* [DenseCap: Fully Convolutional Localization Networks for Dense Captioning | Papers With Code](https://paperswithcode.com/paper/densecap-fully-convolutional-localization#:~:text=The%20architecture%20is%20composed%20of%20a)

Bottom-Up and Top-Down Attention Model:

* [bottom-up-attention/README.md at master · peteanderson80/bottom-up-attention (github.com)](https://github.com/peteanderson80/bottom-up-attention/blob/master/README.md)
* [[1707.07998] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering (arxiv.org)](https://arxiv.org/abs/1707.07998#:~:text=In%20this%20work,%20we%20propose%20a)
* [1707.07998 (arxiv.org)](https://arxiv.org/pdf/1707.07998)