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

DenseCap:

* [jcjohnson/densecap: Dense image captioning in Torch (github.com)](https://github.com/jcjohnson/densecap)

Các bước thực hiện chính của model:

* Trích xuất vùng ảnh (Region Proposal Network): RPN quét toàn bộ ảnh tạo ra các đề xuất về những vùng có thể chứa các đối tượng và khu vực cần mô tả. Tạo ra một số lượng lớn các bounding box dựa trên các anchor points với nhiều tỷ lệ và kích thước khác nhau.
* Trích xuất đặc trưng (Feature Extraction): Sau khi xác định các vùng đề xuất, DenseCap sử dụng một mô hình CNN để trích xuất đặc trưng của từng vùng này. Mỗi vùng đề xuất được đưa qua một CNN để thu được các vector đặc trưng đại diện. Các vector đặc trưng này bao gồm thông tin về cấu trúc và ngữ cảnh của vùng ảnh đó.
* Tạo mô tả (Captioning): Với mỗi vùng ảnh được phát hiện và trích xuất đặc trưng, mô hình sử dụng một Captioning Network để tạo ra mô tả.Mỗi vector đặc trưng của vùng ảnh sẽ được chuyển qua một Recurrent Neural Network (RNN) hoặc Long Short-Term Memory (LSTM) để tạo ra chuỗi mô tả.