Extracting Large-Scale Multimodal Datasets From Web Archives

Master's Thesis Presentation

Thilo Brummerloh

October 21 2024

Leipzig University, Data Science (M.Sc.)

Motivation

- Large-scale ML models, such as LLMs and text-to-image networks, require massive datasets for training.
- Web archives, like Common Crawl, offer rich data but extracting quality image-text pairs is challenging because of their unstructured nature.

Problem Description

- Generating large, high-quality datasets is essential for training text-to-image models.
- Challenge: Extracting meaningful text-image pairs from unstructured web data.

Other Approach

- **LAION-5B**: Uses alt text to extract text-image pairs.
- **Limitations**:
  - Severely limits recall/search space by reducing all possible images to only those that have an alt text.
  - Large image hosting platforms like Flickr, Instagram, and some of the other largest sources from LAION-5B contain AI-generated image alt text.
  - This creates a potential problem when training a text-to-image model due to recursive generation issues.

Alt Text Example - Image Load Failure

- When an image fails to load, the alt text is displayed.
- Enhances accessibility for users with screen readers.

HTML Code Example:

```
<img src="pug.jpg"
alt="Pug looking at the
camera, background is a
courtroom">
```

Figure 1: Image displayed successfully

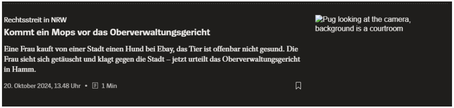


Figure 2: Image missing - Alt text displayed

Limitations of Alt Texts

- **Alt Text Challenges:**
  - Often vague or missing entirely.
  - Common examples include: "USERIMAGE", "IMG_123", which lack meaningful information.
- **Impact:**
  - Low-quality alt text lead to poor image-text associations for training.

Contextual Text Extraction Approach

- **New Approach:**
  - Extract descriptions from the text surrounding images instead of relying on alt texts.
  - Use a fine-tuned BERT model to identify relevant descriptive text.
- **Benefits:**
  - Richer context leads to higher-quality image-text pairs.
  - Enhances model performance in generating or understanding images.

Creating Training Data Using Alt Texts

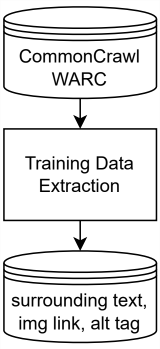
Leveraging Alt Texts for Dataset Creation

- **Objective**: Train a model to identify how image descriptions fit into the context of surrounding text.
- **Text Extraction**:
  - Alt texts are matched in surrounding text.
  - Extract text segments immediately **before and after** each image.
- **Goal**: Build high-quality image-text pairs using context to define good descriptions.
- **Training Signal**: Alt texts in context provide a signal for embedding effective image descriptions.

Creating Training Data Using Alt Texts

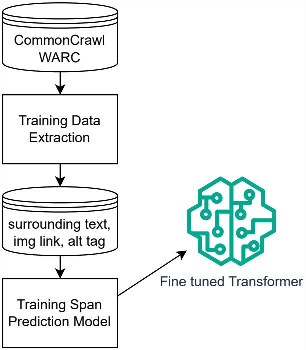
Figure: Illustration of alt text usage and surrounding text extraction for dataset creation.

Approach Overview



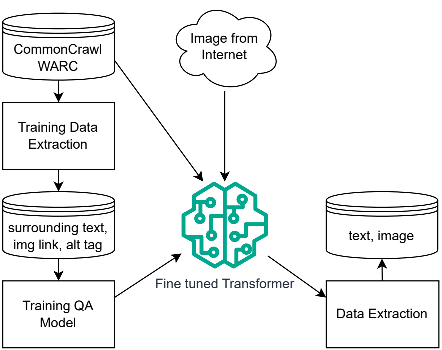
- **Step 1**: Extract alt text-text pairs from WARC archives.

Approach Overview



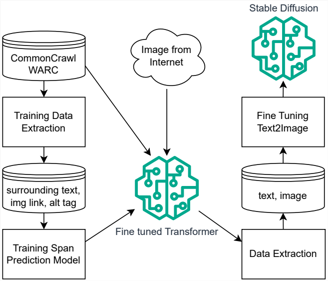
- **Step 1**: Extract alt text-text pairs from WARC archives.
- **Step 2**: Fine-tune BERT for descriptive text extraction.

Approach Overview



- **Step 1**: Extract alt text-text pairs from WARC archives.
- **Step 2**: Fine-tune BERT for descriptive text extraction.
- **Step 3**: Use the fine-tuned model to extract text-image pairs and validate with CLIP.

Approach Overview



- **Step 1**: Extract alt text-text pairs from WARC archives.
- **Step 2**: Fine-tune BERT for descriptive text extraction.
- **Step 3**: Use the fine-tuned model to extract text-image pairs and validate with CLIP.
- **Step 4**: Fine-tune Stable Diffusion using validated pairs.

Step 1 - Extracting Training Data

- Extract text, alt texts, and image links from HTML within WARC files.
- Aim: Identify potential descriptive text-image pairs.
- Processing large-scale data from Common Crawl archives.
- Example: Searching for occurrences of alt attributes in the HTML text beyond just alt texts.

Step 2 - Fine-Tuning BERT for Text Extraction

- Fine-tune a BERT model on spans of text around images.
- Focus: This is a text segmentation task to identify relevant text spans.
- **Loss Functions**: Use of Sparse Categorical Crossentropy (SCCE) and Soft IoU Loss for optimizing start and end token prediction.
- Metrics such as Exact Match and Intersection-over-Union are calculated and logged

Loss Functions for Fine-Tuning BERT

- **Sparse Categorical Crossentropy (SCCE)**
  - Measures prediction accuracy for start and end tokens.
  - Suitable for classification over tokens.
- **Soft Intersection over Union (IoU) Loss**
  - Measures overlap between predicted and true spans.
  - Focuses on improving the quality of span predictions.

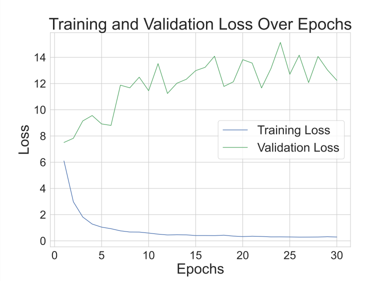
Step 3 - Extracting Text-Image Examples

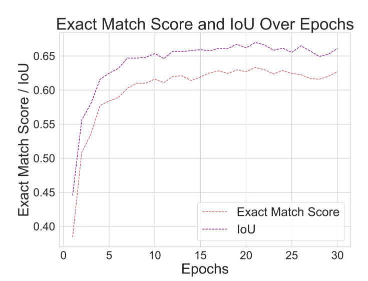
- Use the fine-tuned BERT model to extract descriptive text spans for images.
- Apply CLIP scores to validate the alignment between text and images.
- Only retain pairs with high semantic alignment, ensuring quality of description.

Step 4 - Fine-Tuning Stable Diffusion

- Use validated text-image pairs to fine-tune Stable Diffusion.
- Fine-tuning enhances the model's performance in generating images from complex prompts.
- Evaluation shows improved image quality and fidelity to descriptions.

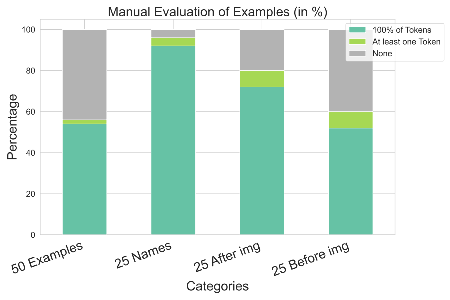
Quality of Model - Training Loss and IoU Performance





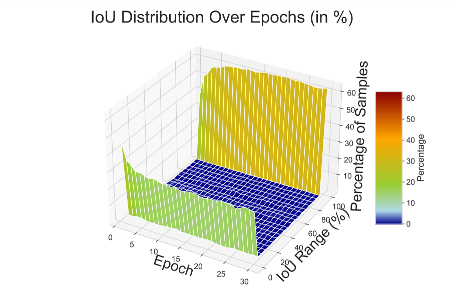
- Training loss decreased over time, while validation loss increased.
- The IoU score improved over training epochs, showing better overlap between predicted spans and actual descriptions.
- Acknowledge overfitting possibility due to increased validation loss despite improving IoU.

Quality of Model - Description Types Evaluation



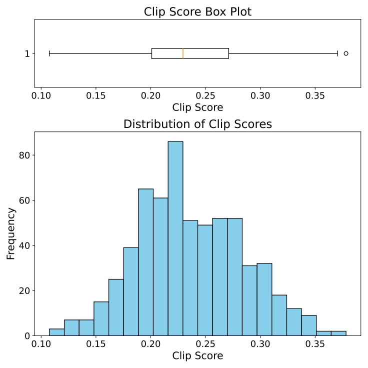
- Evaluated the model's ability to recognize different types of image descriptions.
- Results highlight variations in model performance based on description types.

Quality of Model - IoU Distribution



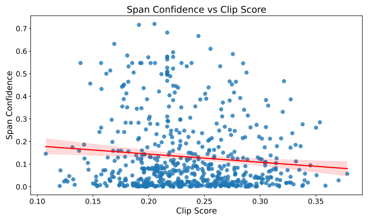
- Samples are either predicted completely or not at all.

Quality of Text-Image Dataset - CLIP Score Analysis



- CLIP scores indicate the alignment quality between text and images.
- Cutoff for LAION-5B was 0.28 which removed 90% of pairs.

Quality of Text-Image Dataset - Confidence Analysis



- Span confidence analysis shows the model's confidence in the extracted descriptions.
- Higher confidence scores correlate with better text-image alignment.

Evaluation Dataset Creation for Fine-Tuned Model

Prompt Generation and Image Creation

- **Objective**: Evaluate the performance of the fine-tuned Stable Diffusion model by comparing generated images to a reference.
- **Prompt Design**:
  - **Two Prompt Sets**:
    - **Nonsensical Prompts**: 100 groups, 5 prompts each, designed to create abstract or random scenarios.
    - **Sensible Prompts**: 100 groups, 5 prompts each, designed to describe meaningful objects or scenes.
- **Prompt Complexity**: Progressively detailed prompts to test model adaptation from simple to complex descriptions.

Evaluation Dataset Creation for Fine-Tuned Model (cont.)

- **Image Creation**:
  - **Stable Diffusion Versions**:
    - **Original Model**: Generated 1,000 images based on the two prompt sets.
    - **Fine-Tuned Model**: Generated 1,000 images using the same prompt sets.
- **Evaluation**:
  - **Manual Evaluation**:
    - **Conformity Assessment**: Rated image quality on five aspects, with scores ranging from 0 to 5.
  - **Automatic Evaluation**:
    - **CLIP Score Analysis**: Calculated to assess semantic alignment between generated images and prompts.
    - **Compliance and Complexity Metrics**: Used to evaluate how well the images matched prompt requirements and their diversity.

Comparison of Generated Images Using Sensible Prompts

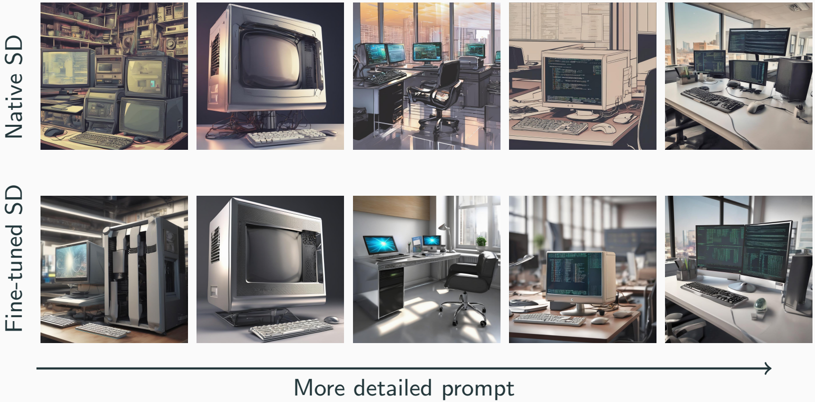
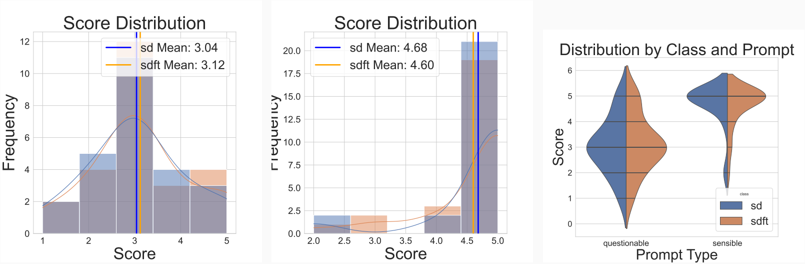


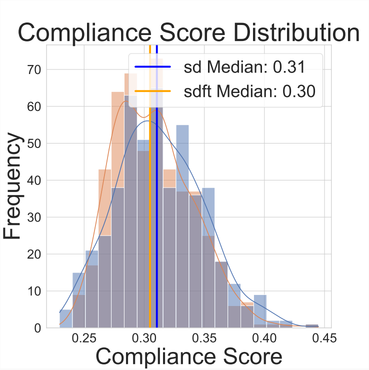
Figure 3: Images generated with SD and SDFT using prompts: "computer" to "A shiny computer in an office displaying code running a simulation"

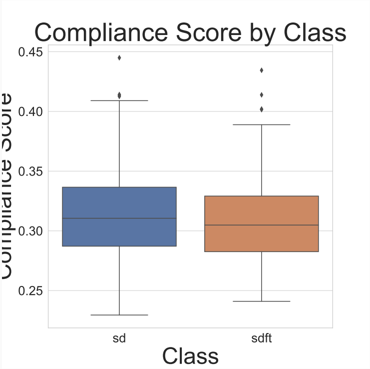
Manual Evaluation of Fine-Tuned SDXL - Sensible vs. Questionable Prompts



- Comparison of performance between fine tuned and native model.
- Nonsensical Prompts were harder for the models to generate.

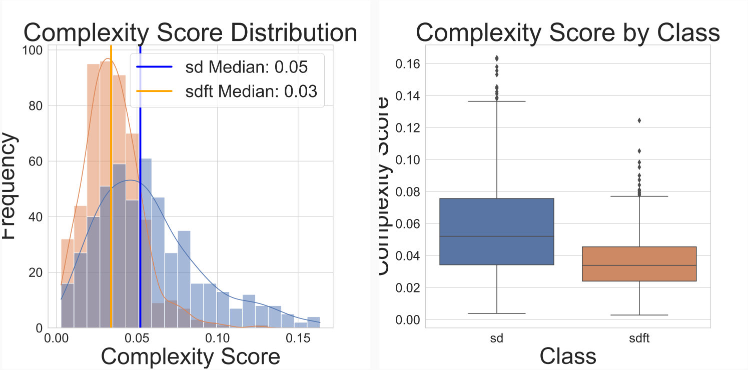
Automatic Evaluation - Compliance Analysis





- Compliance scores measure the degree to which generated images match their textual descriptions.

Automatic Evaluation - Complexity Analysis



- Complexity scores reflect the diversity and richness of descriptions in the dataset.

Conclusion - Contributions

- Developed a scalable pipeline for extracting multimodal datasets.
- Fine-tuned models for better text-image association beyond basic alt text extraction.
- Developed pipeline to improve text-to-image generation capabilities through fine-tuning of Stable Diffusion with automatic evaluation.

Conclusion - Challenges

- Encountered challenges with noisy and incomplete data from web archives.
- Difficulty in ensuring consistent image-description alignment.

Conclusion - Fixes

- Improve model quality to make the downstream model training more useful.
- Using an easier method to manage datasets, possibly in SQLite

Conclusion - Future Work

- Extend the pipeline to extract even larger datasets with improved text-image alignment.
- Explore applications beyond images, such as video or audio descriptions.