WikiWeb2M: A Page-Level Multimodal Wikipedia Dataset

Andrea Burns
Boston University*

Krishna Srinivasan
Google

Joshua Ainslie
Google

Geoff Brown
Google

Kate Saenko
FAIR, Boston University

Bryan A. Plummer
Boston University

Jianmo Ni
Google

Mandy Guo
Google

Abstract

Webpages have been a rich resource for language and vision-language tasks. Yet only pieces of webpages are kept: image-caption pairs, long text articles, or raw HTML, never all in one place. Webpage tasks have resultingly received little attention and structured image-text data underused. To study multimodal webpage understanding, we introduce the Wikipedia Webpage 2M (WikiWeb2M)[0] suite; the first to retain the full set of images, text, and structure data available in a page. WikiWeb2M can be used for tasks like page description generation, section summarization, and contextual image captioning.

Keywords: Multimodal Data, Webpages, Machine Learning, Text Generation, Vision and Language

Introduction

Webpages are multimodal, structured content which can been used for pretraining and fine-tuning. Large scale noisy datasets scraped from the web have been used to pretrain large language or contrastive models (Raffel et al., 2020; Jia et al., 2021). Downstream tasks built from webpages have included instruction following, image captioning, news captioning, image-sentence retrieval, and image-article retrieval (Gur et al., 2022; Biten et al., 2019; Tan et al., 2022). Yet little prior work has studied tasks to evaluate multimodal webpage understanding itself.

Many classification and generation problems could be studied with webpages: taxonomic webpage classification, webpage retrieval, web image captioning, and webpage summarization. However, to date there is no open source, multimodal dataset that retains all webpage content. *E.g.*, the Wikipedia Image Text (WIT) dataset (Srinivasan et al., 2021) does not keep HTML structure and misses out on many text sections, as shown in Table 1. Unified text, image, and structure data would allow for greater study of multimodal content understanding with many-to-many text and image relationships. As a result, we propose the new Wikipedia Webpage (WikiWeb2M) dataset of over 2M pages, which unifies webpage content to include all text, images, and their location (*e.g.*, section index) in one example. Table 2 (left) includes the number of pages, sections, and images, along with sample counts for downstream tasks.

Figure 1 (left) shows how one webpage can be used for page description, section summarization, and contextual captioning. These tasks can improve interaction with web content, *e.g.*, a page description may provide a user who is blind more agency by allowing them to preview content before listening to the entire body with a screen reader (Vtyurina et al., 2019). On top of aiding assistive technology, tasks like contextual image captioning and section summarization can be used for modern content generation, as there is growing interest in providing multimodal snippets from the web (Nkemelu et al., 2023).

The WikiWeb2M Dataset

WikiWeb2M is created by rescraping the ∼2M English articles in WIT. Each webpage sample includes the page URL and title, section titles, text, and indices, images and their captions, and more; see Figure 1 (right). This differs from WIT which defined individual samples as image-caption pairs with additional metadata (*e.g.*, originating section title).

We shuffle the WIT webpages to define a random 1.8M/100K/100K train/val/test split. Table 2 (left) shows the number of pages, sections, and images in our dataset after additional processing. In particular, we only retain content sections (*e.g.*, not the "See Also" section). For images, we keep JPEG and PNG and require the dimensions be greater than 1px to allow for a greater diversity of images to be included (*e.g.*, icons)[1]. We include metadata on image dimensions to allow for additional filtering.

In Table 1, we report the number of sections and images compared to the English subset of WIT. We

*Work was done during an internship at Google.

[0]Data is readily available at https://github.com/google-research-datasets/wit/blob/main/wikiweb2m.md

[1]We release image URLs, where they can be fetched.

add nearly 1M total images to the dataset by keeping the images on a webpage regardless of whether they have image captions. We break down section counts by type: structural, heading, text, image, and both text and image. Structural and heading sections do not contain immediate section text (the former have subsections). For heading sections, the section content either linked to a different article, was empty, or only had tables. A notable 6.8M text sections are in WikiWeb2M, none of which were available in WIT.

The WikiWeb2M Tasks

We now describe WikiWeb2M's suite of multimodal generation tasks and task data processing. Table 2 (left) shows data statistics and (right) downstream task performance when using T5 and ViT base models (Raffel et al., 2020; Dosovitskiy et al., 2021).

Page Description Generation The goal is to generate a description of a page given the rest of the webpage's image, text, and structure. We use the Wikipedia-provided page descriptions for each article. We retain a page if the description has at least five words. A small subset of Wikipedia pages are lists[2]; we remove pages that explicitly have "list_of" in their URL or fewer than two rich content sections.

Section Summarization The goal is to generate a sentence that highlights the section's content given images and (non-summary) text in the section and other context sections. We take advantage of the leading sentence bias and use the first sentence of a section its pseudo summary. In a small pilot, a majority of human annotators also deemed the first sentence as a reasonable summary. A section serves as a target section if it has at least five sentences, contains neither a table nor list, and is not the root section. We filter out the root because the root (first) section is often the page description.

Contextual Image Captioning (Nguyen et al., 2022) proposed Wikipedia image captioning given the image's webpage context. With WikiWeb2M, we can now utilize the entire webpage context for the image instead of just the section it originally came from. We only allow target images to be those from WIT to ensure quality captions. Following prior work, we also use the reference description as the ground truth caption to be generated and require it must have at least three words. But, we do not input the attribution description, as it often contains large overlap with the reference description.

Results Table 2 (right) shows results for each task. For contextual image captioning and section summarization we verify that WikiWeb2M's additional sections (compared to only inputting the target section for image captioning or summarization) improve task performance; page description generation is only made possible with our dataset.

[2]For example, `https://en.wikipedia.org/wiki/List_of_mammals_of_the_United_States`

References

[Biten et al.2019] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *CVPR*.

[Dosovitskiy et al.2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[Gur et al.2022] Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2022. Understanding html with large language models.

[Jia et al.2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

[Nguyen et al.2022] Khanh Nguyen, Ali Furkan Biten, Andres Mafla, Lluis Gomez, and Dimosthenis Karatzas. 2022. Show, interpret and tell: Entity-aware contextualised image captioning in wikipedia.

[Nkemelu et al.2023] Daniel Nkemelu, Peggy Chi, Daniel Castro Chin, Krishna Srinivasan, and Irfan Essa. 2023. Automatic multi-path web story creation from a structural article.

[Raffel et al.2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

[Srinivasan et al.2021] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR*.

[Tan et al.2022] Reuben Tan, Bryan A. Plummer, Kate Saenko, J. P. Lewis, Avneesh Sud, and Thomas Leung. 2022. Newsstories: Illustrating articles with visual summaries. In *ECCV*.

[Vtyurina et al.2019] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. Bridging screen readers and voice assistants for enhanced eyes-free web search. In *ASSETS*.