WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning

Krishna Srinivasan
Google
krishnaps@google.com

Karthik Raman
Google
karthikraman@google.com

Jiecao Chen
Google
chenjiecao@google.com

Michael Bendersky
Google
bemike@google.com

Marc Najork
Google
najork@google.com

ABSTRACT

The milestone improvements brought about by deep representation learning and pre-training techniques have led to large performance gains across downstream NLP, IR and Vision tasks. Multimodal modeling techniques aim to leverage large high-quality visio-linguistic datasets for learning complementary information across image and text modalities. In this paper, we introduce the Wikipedia-based Image Text (WIT) Dataset to better facilitate multimodal, multilingual learning. WIT is composed of a curated set of 37.5 million entity rich image-text examples with 11.5 million unique images across 108 Wikipedia languages. Its size enables WIT to be used as a pretraining dataset for multimodal models, as we show when applied to downstream tasks such as image-text retrieval. WIT has four main and unique advantages. First, WIT is the largest multimodal dataset by the number of image-text examples by 3x (at the time of writing). Second, WIT is massively multilingual (first of its kind) with coverage over 100+ languages (each of which has at least 12K examples) and provides cross-lingual texts for many images. Third, WIT represents a more diverse set of concepts and real world entities relative to what previous datasets cover. Lastly, WIT provides a very challenging real-world test set, as we empirically illustrate using an image-text retrieval task as an example. WIT Dataset is available for download and use via a Creative Commons license here: https://github.com/google-research-datasets/wit.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**; **Multimedia and multimodal retrieval**; **Image search**; *Structure and multilingual text search*; • **Computing methodologies** → **Machine learning**.

KEYWORDS

machine learning; neural networks; multimodal; multilingual; image-text retrieval; wikipedia; dataset;

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07.
https://doi.org/10.1145/3404835.3463257

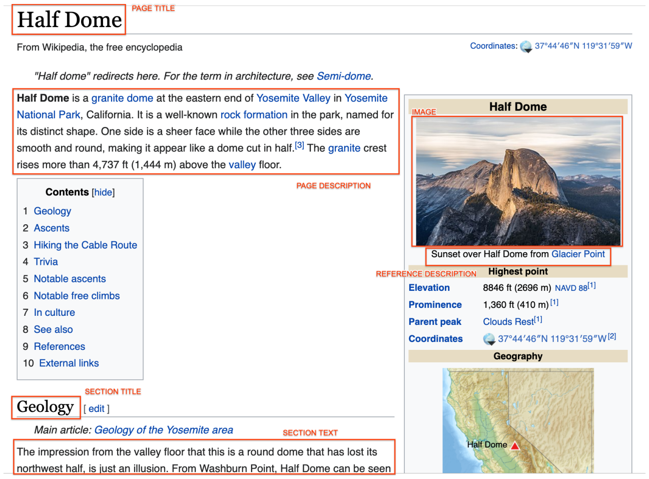
ACM Reference Format:
Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463257

1 INTRODUCTION

Deep learning has fundamentally revolutionized the fields of NLP, IR and Vision via our ability to have a rich semantic understanding of texts and images. Notable examples of this include Deep CNN models [30, 35] which set the bar for standard vision tasks like image recognition and image classification. Attention based transformer models [36] like BERT [9] have likewise enabled achieving new benchmark performance across a myriad of text understanding / NLP / IR tasks. These transformational advances have also found their way to multimodal tasks such as image-text retrieval / search [15] and image captioning [37, 42]. Multimodal models – such as ViLBERT [23], UNITER [7], Unicoder-VL [18] amongst others [1, 20, 33] – are able to jointly model the complex relationships between text and visual inputs leading to wins in downstream tasks like image search, Visual Question Answering (VQA) [2] and Visual Commonsense Reasoning (VCR) [41].

Accompanying the modeling improvements across these advancements, an equally critical aspect is the leveraging of massive datasets to enrich representation learning – often via unsupervised *pretraining*. Increasingly, the efficacy of a model correlates strongly with the size and quality of pretraining data used. For instance, cutting-edge language models like BERT [9] and T5 [28] rely on increasingly larger text datasets spanning from those in the O(100M) range like Wikipedia, BooksCorpus [44] to datasets with billions of examples like C4 [28] and mC4 [39]. Similarly, vision models [10] are reliant on large corpora, such as ImageNet-21k [8] – which with 14M images is among the largest public datasets. This scale is important since studies have shown performance increases logarithmically with dataset size [34]. Another key dimension of language datasets is the number of languages covered. By transitioning from English-only to highly multilingual language datasets, models like mT5 [39] and mBERT [38], are an important step for researchers driving globally, equitable availability of information.

Multimodal visio-linguistic models are no different, and rely on a rich dataset to help them learn to model the relationship between images and texts. However as seen in Table 1, the scale of current



Half Dome

From Wikipedia, the free encyclopedia

Coordinates: 37°44′46″N 119°31′59″W

"Half dome" redirects here. For the term in architecture, see Semi-dome.

Half Dome is a granite dome at the eastern end of Yosemite Valley in Yosemite National Park, California. It is a well-known rock formation in the park, named for its distinct shape. One side is a sheer face while the other three sides are smooth and round, making it appear like a dome cut in half.[3] The granite crest rises more than 4,737 ft (1,444 m) above the valley floor.

Contents [hide]

1 Geology
2 Ascents
3 Hiking the Cable Route
4 Trivia
5 Notable ascents
6 Notable free climbs
7 In culture
8 See also
9 References
10 External links

Geology [edit]

Main article: Geology of the Yosemite area

The impression from the valley floor that this is a round dome that has lost its northwest half, is just an illusion. From Washburn Point, Half Dome can be seen

Half Dome

Sunset over Half Dome from Glacier Point

Highest point

Elevation	8846 ft (2696 m) NAVD 88[1]
Prominence	1,360 ft (410 m)[1]
Parent peak	Clouds Rest[1]
Coordinates	37°44′46″N 119°31′59″W[2]

Geography

Figure 1: The Wikipedia page for Half Dome, Yosemite, California via Wikimedia Commons with examples of the different fields extracted and provided in WIT.

public datasets pales in comparison to image-only or text-only ones, with the 30K-sized Flickr [40] and 3.3M-sized Conceptual Captions (CC) [29] being among the largest ones. Having large image-text datasets can significantly improve performance, as a couple of recent works [16, 27] have shown by leveraging larger noisy (proprietary) datasets. Furthermore the lack of language coverage in these existing datasets (which are mostly only in English) also impedes research in the multilingual multimodal space – which we consider a lost opportunity given the potential shown in leveraging images (as a language-agnostic medium) to help improve our multilingual textual understanding [31] or even translate [13].

To address these challenges and advance research on multilingual, multimodal learning we present the Wikipedia-based Image Text (WIT) Dataset. WIT is created by extracting multiple different texts associated with an image (*e.g.,* the reference description seen in Fig 1) from Wikipedia articles and Wikimedia image links. This was accompanied by rigorous filtering to only retain high quality image-text associations. The resulting dataset contains over 37.5 million image-text sets and spans 11.5 million unique images – making WIT the largest multimodal dataset at the time of writing. Furthermore WIT provides unparalleled multilingual coverage – with 12K+ examples in each of 108 languages (53 languages have 100K+ image-text pairs).

It is worth pointing out that by leveraging Wikipedia’s editing, verification and correction mechanism, WIT is able to ensure a high quality bar. In particular, this use of a curated source like Wikipedia contrasts with the approach used to create other existing datasets (*e.g.* CC [29]) which rely on extracting annotations from web crawls. We verified the curated quality of the WIT dataset via an extensive human-annotation process (nearly 4400 image-text examples and 13K judgments across 7 languages), with an overwhelming majority (98.5%) judging the randomly sampled image-text associations favorably.

Empirical results on image-text retrieval tasks (both zero-shot *i.e.,* pretrained model, as well as finetuned model evaluations) demonstrate the potency of the data. The vast richness of Wikipedia texts and images (grounded in a diverse set of real-world entities and attributes) also means that WIT provides for a realistic evaluation set – one that we demonstrate to be challenging for models trained using existing datasets.

Table 1: Existing publicly available image-text datasets pale in comparison to text-only datasets (e.g., mC4 with O(Billions) of examples in 100+ languages) and image-only datasets (e.g., 14M in ImageNet-21k).

Dataset	Images	Text	Languages
Flickr30K [40]	32K	158K	< 8
SBU Captions [25]	~1M	~1M	1
MS-COCO [22]	~330K	~1.5M	< 4
CC [6]	~3.3M	~3.3M	1
WIT	**11.5M**	**37.5M**	**108**

2 RELATED WORK

Visio-Linguistic (VL) datasets: Flickr30K [40] was among the first datasets that helped drive early research in this space. Similar to other such early datasets (*e.g.* the 330k example MS-COCO), it was created by having crowd sourced (Mechanical Turk) workers provide captions for ~30K images (sampled from Flickr). While the explicit human-based captioning helps ensure quality, the resulting datasets have been recognized as insufficient for significant real-world improvements given that they are small and expensive to construct [11, 43]. Furthermore, this manual effort has meant extending to other languages has proven to be quite challenging. Consequently there exists only a handful of non-English data collections such as Multi30K-DE (German) [12], DeCOCO (German) [14], Multi30K-FR (French) [11], Multi30K-CS (Czech) [4], YJCaptions26k (Japanese) [24], MLM Dataset (based on Wikipedia in 3 languages) [3] and MS-COCO-CN (Chinese) [21].

An alternative paradigm to creating such datasets is demonstrated by the Conceptual Captions (CC) dataset [29]. By leveraging the alt-text annotations for images from a web crawl, the resulting dataset was significantly larger than previous ones (~3.3M image-text pairs). The drawback with this approach is the reliance on complex filtering rules and systems to ensure data quality. Unfortunately this makes these *extraction*-based datasets – like CC and the recently proposed CC12M [6] – hard to extend and significantly impacts their coverage and diversity. Perhaps unsurprisingly, the complex filtering logic has meant that this approach has so far only been successfully applied to curate English data collections.

WIT looks to achieve the best of both worlds by leveraging an extractive approach on a clean, curated multilingual repository of human knowledge with its accompanying images, illustrations and detailed text descriptions (Wikipedia).

VL models: A slew of models have been proposed to leverage the above datasets (either for unsupervised pretraining or finetuning). For example, ViLBERT [23] uses MS-COCO and CC for pretraining a multimodal transformer based model. UNITER [7] leverages these datasets and pretrains on tasks like image-text matching and word region alignment. Similarly, models like VL-BERT [33], VisualBERT [20], ImageBERT [26], B2T2 [1] and Unicoder-VL [19], all pretrain on CC or similar datasets using a variety of objectives and tasks. Efficacy of these models is often studied on downstream tasks

Table 2: Example of texts extracted for Half Dome example

Field Name	**Text**
Page Title	Half Dome, Yosemite
Canonical Page URL	en.wikipedia.org/wiki/Half_Dome
Page Description	Half Dome is a granite dome at the eastern end of Yosemite Valley in Yosemite National Park, California. It is a well-known rock formation ...
Reference Description	Sunset over Half Dome from Glacier Point
Attribution Description	English: Half Dome as viewed from Glacier Point, Yosemite National Park, California, United States.

Table 3: Statistics of the final WIT dataset and availability of different fields. Tuple refers to one entry in the dataset comprising the image, the three different possible texts and the context. Context texts include the page and (hierarchical) section titles and their respective descriptions

Type	Train	Val	Test	**Total / Unique**
Rows / Tuples	37.04M	261K	210.1K	**37.5M**
Unique Images	11.4M	58K	56.9K	**11.5M**
Ref. Text	16.9M	149.8K	104K	**17.1M / 16.6M**
Attr. Text	34.7M	192.6K	199.7K	**35.1M / 10.9M**
Alt Text	5.3M	29K	29K	**5.4M / 5.2M**
Context Texts	-	-	-	**119.4M**

like image-text retrieval, referring expressions, image captioning, *e*tc using Flickr30K, MS-COCO and similar curated collections. These models have also shown that a larger and more varied data collection, improves downstream task performance.

3 WIT: WIKIPEDIA IMAGE TEXT DATASET

We would like to marry the benefits of curated datasets like Flickr30K and MS-COCO (consistent, high quality image text pairs) with those of extractive datasets like CC (automatically created and scalable), while also creating a multilingual and heterogeneous dataset. To do so, we leverage Wikipedia, which inherently uses crowd-sourcing in the data creation process – via its editorial review process – to ensure quality, freshness and accuracy of content. However, even Wikipedia extractions cannot be directly used as is, due to a plethora of low-information (e.g., generic) image-text associations which would not help VL learning. In the remainder of this section, we describe the WIT creation process and detail the filtering processes we introduced to ensure that only the most useful data is selected.

3.1 Wikipedia Crawl Data

We started with all Wikipedia content pages (*i.e.,* ignoring other pages that have discussions, comments and such). These number about ~124M pages across 279 languages. We used a Flume [5] pipeline to programatically process, filter, clean and store the Wikipedia data. We next extracted images and different texts related to the image along with some contextual metadata (such as the page URL, the page title, description . . .). This yielded about ~150M tuples of *(image data, texts data, contextual data)*, which were the input to the different filters described in the subsequent sections.

Note that there tends to be a wide variance of HTML formatting / layouts used for image captions across (and sometimes even within) Wikipedias in different languages, and hence our extraction rules needed to be particularly robust to ensure high coverage.

3.2 The Texts used in WIT

The texts describing the images come from multiple different sources. The three *directly* associated with the image are:

(1) **Reference description** (abbreviated as *ref*): This is the caption that is visible on the wiki page directly below the image. This is the least common among the three (present in ~24M of the tuples) but tends to be the most topical and relevant.
(2) **Attribution description** (abbreviated as *attr*): This is the text found on the Wikimedia page of the image. This text is common to all occurrences of that image across all Wikipedias and thus can be in a language different to the original page article. Often this text is multilingual *i.e.,* with image descriptions in multiple languages. 138M+ of the 150M tuples have this field – though the vast majority of these are uninformative or noisy. However the remaining have rich semantic descriptions of the images that we would like to extract.
(3) **Alt-text description** (abbreviated as *alt*): This is the "alt" text associated with the image. While not visible in general, it is commonly used for accessibility / screen readers. Despite this (surprisingly) we discovered that this was the least informative of the three texts and in most cases was simply the image file name (We found that of the 121M+ tuples containing this text, only a small fraction to be meaningful descriptions of the image).

In addition to these, we also note that the context part of the tuple contains additional texts indirectly associated with the image (such as the section text or page title). A complete example of these texts, along with other metadata fields we provide and more detailed statistics are available on the WIT dataset Github page.

3.3 Text-based Filtering

To clean the low-information texts, we:

(1) Only retained texts that were at least of length 3.
(2) Removed any alt-text containing generic phrases such as '.png', '.jpg', 'icon' or 'stub' and also phrases with "refer to", "alt text" .. etc.
(3) For attributions and alt-text we enforced that
    - Image is either JPG or PNG (since these texts for other image types were almost always unhelpful). GIF images with a reference description were retained.
    - For tuples without a reference description, we enforced the image is not found in the last sections of the page (*i.e.,* the bibliography, external links and such).

3.4 Image & Image-Text based Filtering

We applied the following filters on the images in the tuples:

(1) To ensure rich images, we required that image height and width were at least 100 pixels.
(2) Based on a detailed analysis, we eliminated images which were either generic or didn't have meaningful text associations. For example, images of maps are prevalent on Wikipedia

Table 4: WIT: Image-Text Stats by Language

Image-Text	# Lang	Uniq. Images	# Lang
total > 1M	9	images > 1M	6
total > 500K	10	images > 500K	12
total > 100K	36	images > 100K	35
total > 50K	15	images > 50K	17
total > 12K	38	images > 12K	38

for denoting locations. However since these are generic and not specific to the actual location, the text association is often incorrect and hence we removed them. Other such noise patterns included common images (*e.g.,* tiny icons), placeholder images and generic *missing images*.

(3) We only retained images that have a research-permissive license such as Creative Commons (the text of Wikipedia is licensed under a CC-BY-SA license).

(4) Lastly we found that certain image-text pairs occurred very frequently. These were often generic images that did not have much to do with the main article page. Common examples included flags, logos, maps, insignia and such. To prevent biasing the data, we heavily under-sampled all such images.

3.5 Additional Filtering

To ensure a high-quality dataset free of inappropriate content, we removed tuples with questionable images or texts as done by previous works [29]. In particular we aimed to remove pornographic / profane / violent / . . . content using multiple techniques based on sophisticated image understanding and multilingual text understanding models. Overall these filters help improve data quality while only eliminating < 0.2% of all tuples.

Akin to other multilingual datasets (*e.g.,* mC4 [39]), we restricted our initial version to only the top 100 languages and hence only retained tuples for languages with 12K+ tuples. Lastly we created partitioned the data into training, validation and test splits (with 50K images for the latter two) by ensuring that each image only occurs in a single split.

3.6 Analyzing the WIT Data

As seen in Table 1, the resulting dataset is significantly larger than previous ones with over 37M (image, text(s), context) tuples, spanning 108 languages and covering 11.5 million unique images. Among its many unique aspects and firsts:

- **Multiple texts per image**: WIT provides for multiple different kinds of texts per image. More than half of the tuples (19.4M) have two or more of reference, attribution and alt-texts. Table 3 provides some more detailed statistics of the coverage of the different texts. Overall with nearly 32M unique image-text pairs, WIT is nearly an order of magnitude larger than prior datasets.
- **Highly multilingual**: As seen in Table 4, WIT has broad multilingual coverage. Nearly half of the 100+ languages contain 100K+ unique image-text tuples and 100K+ unique images.
- **Large cross-lingual coverage**: Images have shown great promise in helping build cross-lingual models [31, 32]. WIT can be used to generate 50M+ cross-lingual pairs (*i.e.,* text descriptions in different languages for the same image) from 3.1M different images using just the reference and alt texts. We expect this number to be even higher when counting attributes, many of which are inherently multilingual.
- **Contextual understanding**: WIT is also the first dataset, providing for understanding image captions *in the context of* the page and surrounding text (incl. ~ $120M$ contextual texts). For the sake of brevity we explore this in future work.

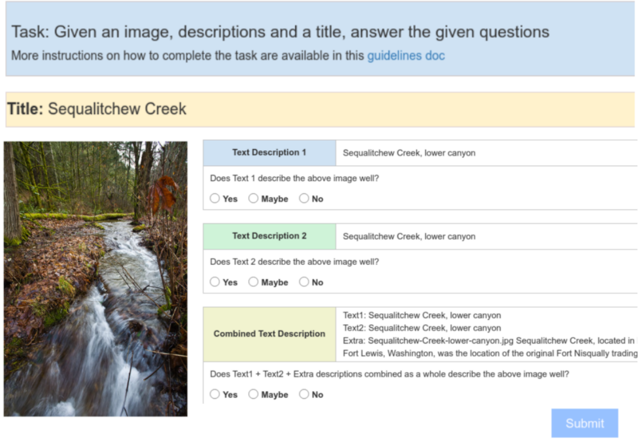
Table 5: Results of the human annotations of data quality. These examples and ratings are included with the dataset.

Text	EN			non-EN		
	%Yes	%Maybe	%No	%Yes	%Maybe	%No
Reference	92.2	4.4	3.3	94.1	2.9	2.9
Attribute	92.2	3.3	4.6	93.1	0.8	6.2
Contextual	98.7	0.7	0.6	96.6	1.8	1.6

3.7 Human Annotator Validation

To further verify the quality of the WIT dataset we performed a study using (crowd-sourced) human annotators. As seen in Fig. 2, we asked raters to answer 3 questions. Given an image and the page title, raters first evaluate the quality of the attribution description and reference description in the first two questions (order randomized). The third question understands the *contextual* quality of these text descriptions given the page description and alt-text description. Each response is on a 3-point scale: "Yes" if the text perfectly describes the image, "Maybe" if it is sufficiently explanatory and "No" if it is irrelevant or the image is inappropriate.

We randomly sampled nearly 4.4k examples for this evaluation. To maximize rating quality we used a language identification filter on the attribution to show raters examples in the language of their expertise. In addition to rating ~ $3k$ examples in English, we also rated 300 examples in German, French, Spanish, Russian, Chinese and 100 examples for Hindi. (We chose these languages to capture different language families and different sizes – Hindi is only 65^{th} in size). Each example was rated by three raters and majority label was used (Maybe being selected if no majority). As seen from the results in Table 5, an overwhelming majority of examples were found to be very helpful. The data quality as judged by the raters



Task: Given an image, descriptions and a title, answer the given questions

More instructions on how to complete the task are available in this guidelines doc

Title: Sequalitchew Creek

Text Description 1	Sequalitchew Creek, lower canyon

Does Text 1 describe the above image well?

○ Yes ○ Maybe ○ No

Text Description 2	Sequalitchew Creek, lower canyon

Does Text 2 describe the above image well?

○ Yes ○ Maybe ○ No

Combined Text Description	Text1: Sequalitchew Creek, lower canyon Text2: Sequalitchew Creek, lower canyon Extra: Sequalitchew-Creek-lower-canyon.jpg Sequalitchew Creek, located in Fort Lewis, Washington, was the location of the original Fort Nisqually trading

Does Text1 + Text2 + Extra descriptions combined as a whole describe the above image well?

○ Yes ○ Maybe ○ No

Figure 2: Human Annotation Template Example

Table 6: Zero-shot evaluation for models using different text fields on WIT Image-Text Retrieval test sets

Pretrain setup		WIT-All		WIT-EN		WIT-I18N	
Data	**Text**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**
WIT	**ref**	0.126	0.258	0.169	0.358	0.114	0.236
WIT	**attr**	0.293	0.55	0.272	0.523	0.293	0.523
WIT	**ref+attr**	**0.346**	**0.642**	**0.344**	**0.64**	**0.344**	**0.633**
CC	**text**	0.048	0.122	0.072	0.186	0.041	0.11

compares favorably to similar evaluations of existing datasets [29]. Both reference and attribution were found to be high-quality (with a slight edge to reference description). The responses to the third question (which provides the page context) also validated our hypothesis that the relevance of image captions is influenced by the context as seen by the near-perfect ratings when considering the context. Lastly we found no major difference in performance across the different languages demonstrating the multilingual data quality.

4 MULTIMODAL EXPERIMENTS WITH WIT

In this section, we empirically demonstrate the efficacy of the WIT dataset both as a pretraining dataset as well as an evaluation set for a new image-text retrieval task.

4.1 Experiment Details

Model: For this analyses, we leveraged a two-tower or *dual-encoder* model, inspired by previous works that used them to learn multilingual, multimodal models [31]. As the name suggests, the model has two encoders – one to encoder the text and the other to represent the images. While the text input to the model was a bag of words, the image tower, the image was first embedded in a manner similar to [17]. The final embeddings of these two towers is then combined using their cosine similarity, which in turn is optimized using a batch softmax loss. Specifically, for a batch of n image-text embedding pairs, the complete $n \times n$ similarity matrix is computed (the (i, j) entry being the cosine of the i^{th} image embedding and j^{th} text embedding) and a softmax loss applied on each of the row. Note that only the diagonal entries are considered as positive pairs.

Setup: We used a batch size of 128 for training and a batch size of 1000 for evaluation. The learning rate was set to 5e-7 The optimizer we used was SGD with Momentum. For the text encoder, we used a bag of words model (with ngrams of size 1 and 2). Each ngram was mapped to an a one amongst a million vocabulary buckets using a hash-function to get a 200D embedding. These ngram embeddings were then summed and passed through a simple FFNN and projected to a final 64D embedding, to match the size of the image encoder embedding. The final activation function we used was ReLU.

Evaluation: We evaluated the models on the Flickr30K, Multi30K and MS-COCO test sets, as well as the dedicated test sets released as part of WIT. We also spliced the WIT test sets into English-only and i18n (non-English) to understand any performance differences. In all experiments using WIT for pretraining, we use the entire training set (*i.e.,* data for all languages). We also pretrained a model with Conceptual Captions (CC) dataset to compare against. We used Recall@K (K = 1, 3, 5) as the evaluation metric.

Table 7: Zero-shot Evaluation on Flickr30K, MS-COCO and WIT test sets for Image-Text Retrieval Task

Pretrain	MS-COCO		Flickr30K		WIT-ALL	
	R@1	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**
WIT-ALL	0.074	0.228	0.054	0.165	**0.346**	**0.642**
CC	**0.145**	**0.385**	**0.111**	**0.32**	0.048	0.122

Table 8: Zero-shot Evaluation on Multi30K and WIT I18N test sets (CS, DE, FR) for Image-Text Retrieval Task

Exp	Multi30K-R@5			WIT-R@5		
	CS	**DE**	**FR**	**CS**	**DE**	**FR**
WIT-ALL	**0.006**	**0.005**	**0.006**	**0.553**	**0.562**	**0.599**
CC	0.004	**0.005**	0.004	0.096	0.084	0.104

4.2 Evaluating a zero-shot pretrained model

A common evaluation of image-text datasets is as a pretraining dataset for a model, which is then directly applied to a downstream task – in our case image-text retrieval – without any finetuning (*i.e.,* zero-shot). Since WIT contains multiple different texts associated with an image, we first set about understanding the effect of pretraining models on different fields. As seen in Table 6, the different WIT models all perform quite well on both English and non-English sets. The strongest performance was consistently obtained by the concatenation of reference and attribution descriptions – which we now default to for subsequent experiments. It is worth noting that the model pretrained on CC lags behind those trained on WIT, even on the English-only test set.

To better understand this, we next evaluated the WIT and CC models (in this zero-shot manner) on popular English test collections from Flickr30K and MS-COCO which are more similar to CC. As seen in Table 7, the multilingual WIT model trails the English CC model on these collections, though not as significantly as the gap between WIT and CC on the heldout WIT test sets.

4.3 Understanding multilingual performance

Since WIT encompasses examples from 100+ languages, we next evaluated how multilingual the WIT-based models are. For this, we used Multi30K's three language test sets (Czech (CS), German (DE) and French (FR)). We generated similar language subset datasets from the WIT test set for the same languages (CS, DE, FR) and used that for evaluation. As shown in Table 8, both models struggle on the Multi30K dataset, though again the WIT model shines on the held-out WIT test set. Similar to the Flickr30k dataset, the Multi30k datasets are quite different from the WIT datasets (as we discuss in Sec. 4.5) which may explain this behavior.

4.4 Evaluation On Image/Title Retrieval Task

Lastly, we evaluated on a real-world task that's based on Wikipedia. This retrieval task requires identifying images that can be found on a given Wikipedia page, using only the page title. We ran this evaluation in both a zero-shot setting (*i.e.,* pretrained model directly) and with finetuning on the training set. Unlike the above experiments, here the input to the text encoder was the page title directly. The evaluation was done with the held-out WIT test split using the page title as text. From Table 9, we clearly observe a large performance

Table 9: Zero-shot and Finetuned Evaluation on Wiki (Image, Page Title) test set for Retrieval Task

Exp	Finetuning	WIT-All		
		R@1	R@3	R@5
WIT-EN	None	0.067	0.122	0.152
CC	None	0.012	0.024	0.032
WIT-ALL	WIT-ALL	**0.1**	**0.174**	**0.214**
CC	CC	0.01	0.021	0.029

Table 10: Vocabulary Comparison

Dataset	Unigrams	freq <= 3	pct freq <= 3
CC	149,924	63,800	42.55%
WIT (ref)	867,906	625,100	**72.02%**

Table 11: Language Model Comparison

Dataset A vs B	JSD
Flickr vs Flickr Test	0.1679
COCO vs COCO Test	0.1008
CC vs Flickr Test	0.4844
CC vs COCO Test	0.4746
CC vs WIT	0.3825
WIT vs Flickr Test	**0.6007**
WIT vs COCO Test	**0.5957**

gain on this task using WIT relative to the CC model both with and without finetuning.

4.5 Discussion

The above experiments clearly demonstrated that WIT-based pretrained models perform extremely well (5x+ gains) on the evaluation sets based on Wikipedia data. However, the models do not do as well on other image-text datasets (Flickr30K/Multi30k and MS-COCO). Since the WIT dataset is not lacking in size or diversity, we probed further into what makes these evaluation sets so different from each other.

4.5.1 Vocabulary Analysis. We first analyzed the vocabulary of the two datasets we used for pretraining : WIT and CC. Since Wikipedia is entity heavy with a diverse concept pool, we suspected that the vocabulary of the WIT dataset may reflect this. As shown in Table 10, this was the case with over 72% of WIT unigrams occurring 3 times or less (vs. 43% for CC).

4.5.2 Language Model. This difference is even more stark when compared to the test collections used for evaluation (COCO and Flickr). When we compared the unigram distributions of different data sets using the Jensen-Shannon Divergence (JSD), we found a massive difference in the vocabularies and concept coverage of the data (see Table 11). While the fact that less than a sixth of WIT is English skews these results slightly, the gap between the English-only slice and other datasets remains sizeable.

4.5.3 Image entity Analysis. Part of the reason for this difference is the broad coverage of entities in the WIT dataset. Using an image classification model to tag all WIT images with entities, we found that amongst the ~4.5M entities identified, a large number (≥ 80% *i.e.,* ~3.68M) of the entities occur 3 times or less. Thus similar to the texts, the image data too is very diverse with not much repetition.

4.5.4 Key differences in texts. Text fields in WIT often tend to be descriptive, verbose and use specific terminology. However this causes a mismatch when evaluated on the test collections, which are often terse single line captions of common words and objects. The choice of bag of words likely exacerbates this issue. Perhaps the most important difference is the use of specifics vs general words. As found in the CC work [29], text hypernymization was crucial to creating a dataset closer to those used for evaluation. For example a text like `Two sculptures by artist Duncan McKellar adorn trees outside the derelict Norwich Union offices in Bristol, UK` would be transformed to `sculptures by person adorn trees outside the derelict offices` so as to remove specifics (person names, locations, times etc ..). This is likely the biggest reason why our trained models underperformed on the existing collections. While there are benefits and drawbacks of such hypernymization, we would like to add this in future versions. However there remains significant challenges doing such replacements for a 100+ language dataset consistently and with high quality across languages.

5 FUTURE WORK

In our eagerness and excitement to share the WIT Dataset with the research community, we have just touched the tip of the iceberg by starting out with an image-text retrieval task using a simple dual encoder model. We observed better performance with an ALIGN model [16] which we plan to expand upon in followup work. Given the superior performance of cross-attention multimodal models, WIT can potentially be used in lieu of or in addition to the existing pretraining datasets in models as illustrated by UNITER, Unicoder-VL, VL-BERT, … etc. A range of new i18n tasks can be formulated with WIT as the basis for VQA, VCR and many others. There is also the possibility of using multimodality to enhance multilingual performance, especially for under-represented languages. WIT Dataset provides a crosslingual corpus of text for the same image which could aid in this idea. We also hope to leverage the knowledge base and entities and attributes of WIT to improve Q&A tasks.

6 CONCLUSION

In this paper we introduced the Wikipedia Image Text (WIT) dataset – the largest (at time of writing), multilingual, multimodal, context-rich dataset. By extracting texts associated with images and their surrounding contexts from over a 100 languages, WIT provides for a rich and diverse dataset. As a result, it is well suited for use in a myriad of ways including pretraining multimodal models, finetuning image-text retrieval models or building cross-lingual representations to name a few. Our detailed analysis and quality evaluation, validate that WIT is a high quality dataset with strong image-text alignment. We also empirically demonstrated the use of this dataset as both a pretraining and finetuning set, and in the process uncovered some shortcomings of existing datasets. We believe this can serve as a rich resource to drive research in the multilingual, multimodal space for years to come and enable the community to building better and more robust visio-linguistic models well suited to real world tasks.

REFERENCES

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of Detected Objects in Text for Visual Question Answering. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*. 2131–2140.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*. 2425–2433.

[3] Jason Armitage, Endri Kacupaj, Golsa Tahmasebzadeh, Maria Maleshkova, Ralph Ewerth, and Jens Lehmann. 2020. MLM: A Benchmark Dataset for Multitask Learning with Multiple Languages and Modalities. In *Proc. of the 29th ACM International Conf. on Information & Knowledge Management (CIKM)*. 2967–2974.

[4] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the Third Shared Task on Multimodal Machine Translation. In *Proc. of the Third Conf. on Machine Translation (Volume 2: Shared Task Papers (WMT)*. 304–323.

[5] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R Henry, Robert Bradshaw, and Nathan Weizenbaum. 2010. FlumeJava: Easy, Efficient Data-Parallel Pipelines. *ACM Sigplan Notices* 45, 6 (2010), 363–375.

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. *arXiv preprint arXiv:2102.08981* (2021).

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740* (2019).

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL-HLT)*. 4171–4186.

[10] 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 *Proc. of the 9th International Conf. on Learning Representations (ICLR)*.

[11] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proc. of the 2nd Conf. on Machine Translation (WMT)*. 215–233.

[12] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proc. of the 5th Workshop on Vision and Language (VL)*. 70–74.

[13] John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning Translations via Images with a Massively Multilingual Image Dataset. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. 2566–2576.

[14] Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. 2399–2409.

[15] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv preprint arXiv:2102.05918* (2021).

[17] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-RISE: Graph-Regularized Image Semantic Embedding. *arXiv preprint arXiv:1902.10814* (2019).

[18] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. arXiv e-prints, page. *arXiv preprint arXiv:1908.06066* (2019).

[19] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*. 11336–11344.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557* (2019).

[21] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval. *IEEE Transactions on Multimedia* 21, 9 (2019), 2347–2360.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of the 13th European Conf. on Computer Vision (Part V) (ECCV)*. 740–755.

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proc. of the 33rd Conf. on Neural Information Processing Systems (NeurIPS)*. 13–23.

[24] Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-Lingual Image Caption Generation. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. 1780–1790.

[25] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Proc. of the 25th Annual Conf. on Neural Information Processing Systems (NeurIPS)*. 1143–1151.

[26] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv preprint arXiv:2001.07966* (2020).

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Image* 2 (2021), T2.

[28] 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. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. 2556–2565.

[30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[31] Karan Singhal, Karthik Raman, and Balder ten Cate. 2019. Learning multilingual word embeddings using image-text data. *arXiv preprint arXiv:1905.12260* (2019).

[32] Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proc. of the First Conf. on Machine Translation: Volume 2, Shared Task Papers (WMT)*. 543–553.

[33] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proc. of the 8th International Conf. on Learning Representations (ICLR)*.

[34] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proc. of the 2017 IEEE International Conf. on Computer Vision (ICCV)*. 843–852.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of the 31st Conf. on Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*. 3156–3164.

[38] Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*. 833–844.

[39] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934* (2020).

[40] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[41] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 6713–6724.

[42] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*. 13041–13049.

[43] Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A Visual Attention Grounding Neural Model for Multimodal Machine Translation. In *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 3643–3653.

[44] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proc. of the 2015 IEEE International Conf. on Computer Vision (ICCV)*. 19–27.