WISMIR3
A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches

Florian Schneider and **Chris Biemann**
Language Technology Group, Department of Informatics
Universität Hamburg, Germany
{florian.schneider-1, biemann}@uni-hamburg.de

Abstract

This paper presents WISMIR3, a multi-modal dataset comprising roughly 300K text-image pairs from Wikipedia. With a sophisticated automatic ETL pipeline, we scraped, filtered, and transformed the data so that WISMIR3 intrinsically differs from other popular text-image datasets like COCO and Flickr30k. We prove this difference by comparing various linguistic statistics between the three datasets computed using the pipeline. The primary purpose of WISMIR3 is to use it as a benchmark to challenge state-of-the-art text-image retrieval approaches, which already reach around 90% Recall@5 scores on the mentioned popular datasets. Therefore, we ran several text-image retrieval experiments on our dataset using current models, which show that the models, in fact, perform significantly worse compared to evaluation results on COCO and Flickr30k. In addition, for each text-image pair, we release features computed by Faster-R-CNN and CLIP models. With this, we want to ease and motivate the use of the dataset for other researchers.

1 Introduction

Current multi-modal text-image retrieval approaches already reach over 90% Recall@5 on popular evaluation sets (Wang et al., 2023). The reason for this is definitely due to the advances in visio-linguistic approaches implemented by state-of-the-art models like UNITER (Chen et al., 2020), TERAN (Messina et al., 2021), CLIP (Radford et al., 2021), or BEiT3 (Wang et al., 2023). However, we argue that this is not solely due to the model's architecture but also because of the simplicity of the widely used training data and its similarity to the evaluation data. Although more recent datasets exist, the most popular datasets used to train and evaluate state-of-the-art text-image retrieval methods are still COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). Both datasets comprise short and simple captions created by crowdsourcing workers for Flickr images showing everyday scenes. Schneider et al. (2021) showed that recent multi-modal transformer-based approaches trained on these popular datasets cannot generalize well on out-of-domain data with more complexity and variety. In the mentioned work, two preliminary datasets were introduced. However, during detailed data analysis, we found multiple issues in these preliminary datasets, which we address in this work.

The main contribution of this work is the release of WISMIR3 (**WI**kiCaps **S**ubset for **M**ulti-Modal Text-**I**mage **R**etrieval v**3**)[1], a clean multi-modal dataset, thought of as a benchmark to challenge state-of-the-art text-image retrieval models. WISMIR3 contains more than 300K text-image pairs from Wikipedia, scraped, filtered, transformed, and statistically analyzed by a sophisticated automatic ETL pipeline tool. Further, we provide a detailed overview, discuss and release linguistic statistics of the comprised data, and compare it to COCO and Flickr30K. Additionally, we release pre-computed image features from a popular pre-trained Faster-R-CNN (Ren et al., 2016) model and image and text embeddings from pre-trained CLIP models employing ViT (Dosovitskiy et al., 2021) as the image encoder. With this, we aim to ease the use of the dataset to train, finetune, or evaluate models on the WISMIR3 dataset. By evaluating different state-of-the-art text-image retrieval approaches on WISMIR3 and comparing the results with their performance on COCO and Flickr30k, we show that these models indeed perform much worse on our dataset.

2 Related Work

State-of-the-art approaches for multi-modal text-image retrieval are typically trained on text-image

[1] https://github.com/floschne/wismir3
https://huggingface.co/datasets/floschne/wismir3

pairs. Despite their age, the most popular datasets to train and evaluate models on this task are still COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). COCO is a well-known dataset for various Computer Vision tasks like object detection, object segmentation, image captioning, keypoint detection, human pose estimation, and text-image retrieval. Besides labels and annotations, the dataset contains about 123K carefully selected images from Flickr with five descriptive captions each. Flickr30k contains about 30K icon photographs of everyday activities, events, and scenes from Flickr, where also five different captions describe each image. Both COCO and Flickr30k are datasets designed by researchers and handcrafted by crowdsourcing workers to describe the images with short, simple, and descriptive captions.

Less popular but larger datasets like SBU Captions (Ordonez et al., 2011), Conceptual Captions (Sharma et al., 2018), or Visual Genome (Krishna et al., 2017) are primarily designed for tasks like image-captioning, visual question answering, or visual entailment. However, since they comprise text-image pairs, the datasets are often part of the training data for text-image retrieval approaches. Visual Genome contains about 108K images collected from an intersection of MS COCO and YFCC-100M (Thomee et al., 2016) with captions created by crowdsourcing workers. SBU Caption contains about 1M photos and their captions from Flickr. Conceptual Captions contains approximately 3.3M text-image pairs scraped from billions of websites and automatically transformed and filtered by a sophisticated pipeline.

Further, WIT (Srinivasan et al., 2021) and LAION-5B (Schuhmann et al., 2022) are huge text-image datasets suitable for pre-training vison-language foundation models like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), or BLIP2 (Li et al., 2023). The WIT dataset contains about 37.5M text-image pairs, comprising 11.5M unique images with captions from Wikipedia across 108 different languages. The LAION-5B dataset contains about 5B non-curated text-image pairs scraped from Common Crawl dumps.

Another text-image dataset is WikiCaps (Schamoni et al., 2018), containing about 3.8M text-image pairs from Wikipedia. Captions are taken from the associated Wikimedia image descriptions, mainly in English. This dataset is the basis of WISMIR3 and is of particular interest in this work because the data is from random Wikipedia articles.

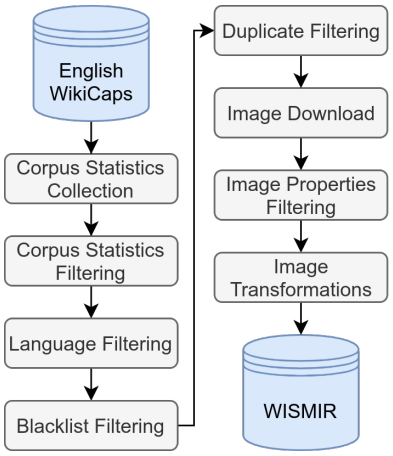


Figure 1: A schematic overview of the pipeline used to collect the WISMIR3 dataset.

Therefore, the captions and images cover a wide range of different topics and concepts.

3 Data Collection Pipeline

A schematic overview of the pipeline used to collect the WISMIR3 dataset, presented by this work, is shown in Figure 1. In the following, more details about the single steps are described.

The input to the pipeline is a CSV file released by the WikiCaps authors, containing 3.8M Wikimedia image file IDs and the corresponding English captions. Since this file format is unhandy to compute statistics or apply transformations, it is converted into a pandas DataFrame, used throughout the whole pipeline.

In the first stage, extensive corpus statistics are collected for each caption using a spaCy pipeline with the "en_core_web_lg" model. These statistics include, for example, the number of tokens and sentences, POS tags of each token, counts of the Universal Dependency tags (Nivre et al., 2020), the language of each sentence, named entities, and ratios between the number of all tokens and nouns or named entities.

The DataFrame is then filtered based on these statistics, as described in the following. Samples are dropped if

- the caption consists of less than 10 or more than 300 tokens
- the caption consists of less than 1 or more than 7 sentences
- the number of tokens in a sentence in the caption is less than 5

- the ratio between all tokens and tokens that are part of named entities does not exceed 0.8

Further samples were removed if the language of every sentence in the caption was not English.

Moreover, since the purpose of this dataset is to challenge text-image retrieval approaches, it is essential that most of the words in an image description are also represented in the image. Hence, we created a blocklist of non-depictable words like "URL", "Sarcasm", "Confusion" and filtered out every sample that contains one or more of these terms.

In the next pipeline stage, the duplicate filtering stage, we remove duplicate captions so that one caption describes at most five different images. This decision was inspired by COCO or Flickr30k, where it is the other way round, i.e., five different captions describe one image.

With the mentioned filtering stages, we reduced the 3.8M WikiCaps samples by about 92% to 304317 samples. After downloading the images, we removed 3431 that were too small or had erroneous data format. We applied the following transformations to every image in the final pipeline stage.

- converting to RGB if it was grayscale before
- resizing while keeping the aspect ratio with bicubic interpolation so that the maximum width and maximum height do not exceed 640 pixels
- compressing to a max of 72 DPI
- converting to and persisting as PNG

The final output of the pipeline is the WISMIR3 dataset, comprising 300886 text-image pairs. A detailed overview is described in the following sections.

4 Dataset Structure and Statistics

4.1 Structure

The textual data of the WISMIR3 is released in two pandas DataFrames[2], one for the training set and one for the test or evaluation set. In addition to the "raw" format, we also release the dataset on HuggingFace[3]. The training and test split comprises 295886 and 5000 randomly chosen text-image pairs, respectively. Besides the caption and the corresponding image filename, both DataFrames contain various linguistic statistics of the caption, as described in Table 1. To compute these statistics, we used spaCy[4] with the "en_core_web_lg" model.

Column Name	Description
wikicaps_id	The row index in the original WikiCaps CSV file
wikimedia_file_id	The Wikimedia File ID of the original image
caption	The caption of the image
tokens	The list of tokens in the caption
num_tok	The number of tokens in the caption
sentence_spans	A list of tuples containing the start and end index of the sentences w.r.t. the list of tokens
num_sents	The number of sentences in the caption
min_sent_len	The minimum length of the sentences in the caption
max_sent_len	The maximum length of the sentences in the caption
num_ne	The number of named entities in the caption
ne_types	A list of the named entity types in the caption
ne_texts	A list of the named entity surface forms in the caption
num_nouns	The number of tokens tagged as NOUN
num_propns	The number of tokens tagged as PROPN
num_conj	The number of tokens tagged as CONJ
num_verb	The number of tokens tagged as VERB
num_sym	The number of tokens tagged as SYM
num_num	The number of tokens tagged as NUM
num_adp	The number of tokens tagged as ADP
num_adj	The number of tokens tagged as ADJ
ratio_ne_tok	The ratio of tokens that belong to named entities versus all tokens of the caption
ratio_noun_tok	The ratio of tokens tagged as NOUN versus all tokens of the caption
ratio_propn_tok	The ratio of tokens tagged as PROPN versus all tokens of the caption
ratio_all_noun_tok	The ratio of tokens tagged as NOUN or PROPN versus all tokens of the caption
image_id	The filename of the image corresponding to this sample
clip_embs_id	The ID of the CLIP image and text embeddings of this sample in the CLIP embeddings tensor
frcnn_embs_id	The filename of the Faster-R-CNN image embedding of this sample

Table 1: The extensive list of the columns and their descriptions contained in WISMIR3.

The images related to the samples are released as single PNG files. Further, we released 36 bounding boxes for regions of interest with corresponding feature vectors extracted by a pretrained Faster-R-CNN (Ren et al., 2016; Yu et al., 2020) model for each image as single NumPy archive files. Additionally, we computed and published the caption and image embedding for each sample computed with two pretrained CLIP (Radford et al., 2021) models employing 16x16 and 32x32 patch ViT (Dosovitskiy et al., 2021), respectively.

Three random samples of WISMIR3, i.e., the images with their corresponding captions, are shown in Figure 2.

4.2 Statistics

In this section, we present a statistical overview of WISMIR3 in Table 2 and, based on this, discuss the contrasts between the dataset and COCO or Flickr30k.

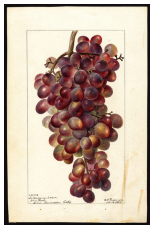
An appreciable difference between WISMIR3, COCO, and Flickr30k becomes apparent when comparing these statistics between the respective datasets. For example, in COCO and Flickr30k, the respective average number of tokens per caption is

[2] https://github.com/floschne/wismir3

[3] https://huggingface.co/floschne/wismir3

[4] https://spacy.io


(a)


(b)


(c)

Figure 2: Randomly chosen images and their captions included in WISMIR3. (a) *Fanta Klassik, 75th anniversay edition of the Fanta soft drink, 2015. Front view of the bottle.* (b) *Image of the Sultanina Rosea variety of grapes (scientific name: "Vitis"), with this specimen originating in Niles, Fremont, Alameda County, California, United States. Source: U.S. Department of Agriculture Pomological Watercolor Collection. Rare and Special Collections, National Agricultural Library, Beltsville, MD 20705.* (c) *"The painting is a design for a poster." image: Three figures dominate the image. A Red Cross nurse stands in the centre. A wounded soldier with a crutch and bandaged head leans on her right arm. On her left a small child in a red dress clings to her skirts; the nurse has her hand resting reassuringly on the child's shoulder. There is the ruin of a building in the background.*

	min	**max**	**avg**
Number of tokens	12	294	59.8
Number of sentences	1	6	2.71
Ratio of NOUN or PROPN tokens	0.0	0.92	0.44
Ratio of named entity tokens	0.0	0.79	0.31
Cosine similarity of caption and image embeddings	0.04	0.53	0.32

Table 2: Various aggregated per-caption statistics in WISMIR3. The cosine similarity was computed using a CLIP model with a ViT using 16x16 patches.

11.34 and 13.49, which is close to the minimum number of tokens and about 4 to 5 times smaller than the average number of tokens per caption in WISMIR3.

Further, by looking at the average ratio of named entity tokens of COCO and Flickr30k, which are 0.02 and 0.03, respectively, it becomes clear that there are almost no named entities in the two datasets. However, in WISMIR3, this ratio lies at 0.44 on average. We argue that in real-world image-retrieval systems, users search for images of specific entities, e.g., with textual queries like "The Eifel Tower at night." instead of general images with queries like "A large iron tower at night". Hence, the training and evaluation data for models powering these real-world systems should contain named entities.

Another difference between WISMIR3 and COCO or Flickr30k is the number of nouns per caption. In COCO and Flickr30k, the average ratio of noun tokens compared to all tokens of a caption is 0.33 and 0.31, respectively, while, in WISMIR3, it is 0.44.

Furthermore, we computed Flesch-Kincaid (Farr et al., 1951) (FK) and Dale-Chall (Chall and Dale, 1995) (DC) readability scores for the captions in the three datasets, which are similar for COCO and Flickr30k but much higher for WISMIR3 (c.f. Figure 3). This suggests a much higher textual complexity of WISMIR3 compared to the two other datasets. That is, COCO and Flickr30k should be easily understood by an average 4th to 6th-grade US student, while WISMIR3 captions are recommended for college students.

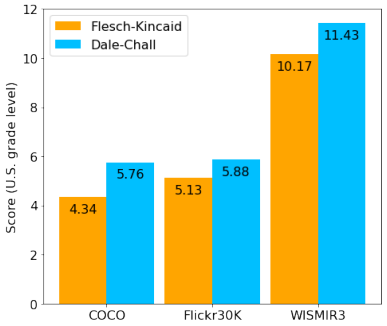


Figure 3: Comparison of Flesch-Kincaid (FK) and Dale-Chall (DC) readability scores of COCO (C), Flickr30k (F), and WISMIR3 (W) captions containing $10^6 \pm 0.1\%$ characters.

We further computed the text-image cosine similarity for each sample in WISMIR3 using a pre-trained CLIP model. With the average similarity of 0.32 being above the minimum threshold of the LAION-400M dataset, we consider the text-image alignment in WISMIR3 as acceptable.

5 Image Retrieval Experiments

This section presents text-image retrieval evaluation results of various recent models on the WISMIR3 dataset and compares them to the models' performances on COCO and Flickr30k. As listed in Table 3, evaluation scores of all listed models on the WISMIR3 (W3) evaluation set are significantly worse compared to the models' performances on COCO (C) and Flickr30k (F30K).

Further observed is that COCO and Flickr30k data did not contribute anything meaningful during TERAN training processes when evaluating the

Text-Image Retrieval (t2i)				
Model	Data	R@1	R@5	R@10
$\text{CLIP}_{\text{ViT-B-16}}$	W3	**47.9**	**72.42**	**80.32**
TERAN_{W3}	W3	15.3	39.6	53.1
$\text{UNITER}_{\text{base}}$	W3	8.76	21.84	29.54
$\text{TERAN}_{\text{COCO}}$	W3	1.1	3.7	5.6
$\text{TERAN}_{\text{F30K}}$	W3	0.9	2.7	4.4
$\text{CLIP}_{\text{ViT-B-16}}$	COCO	**58.4**	**81.5**	**88.1**
$\text{UNITER}_{\text{base}}$	COCO	50.33	78.52	87.16
$\text{TERAN}_{\text{COCO}}$	COCO	42.6	72.5	82.9
$\text{CLIP}_{\text{ViT-B-16}}$	F30K	68.7	90.6	95.2
$\text{UNITER}_{\text{base}}$	F30K	**72.52**	**92.36**	**96.08**
$\text{TERAN}_{\text{F30K}}$	F30K	59.4	84.8	90.5

Table 3: Recall@K evaluation results of different models and evaluation sets on text-image retrieval on the WISMIR3 test set. "W3" stands for WISMIR3. In the model column, the subscript datasets indicate the training data of the TERAN model. For evaluation on COCO, we used the 5k evaluation set. Further, we used CLIP or UNITER in a zero-shot setting without fine-tuning on WISMIR3.

models on WISMIR3. However, one noticeable finding is that the CLIP model[5] performs exceptionally well on WISMIR3 compared to UNITER and even the TERAN model trained on the WISMIR3 training set. Also, UNITER performs much better than TERAN on WISMIR3. Since CLIP was trained on a very large-scale dataset containing more than 400M text-image pairs scraped from random websites, its training data is probably relatively similar to the data contained in WISMIR3 or even comprises the data. Moreover, UNITER was trained on much larger datasets of roughly 5.6M samples compared to WISMIR3.

These findings show that current text-image retrieval approaches perform significantly worse on WISMIR3 than COCO and Flickr30k.

6 Conclusion

This paper presents WISMIR3, a clean multi-modal dataset containing roughly 300K text-image pairs. The dataset comprises images with corresponding captions from Wikipedia using WikiCaps as the source dataset. By implementing a sophisticated automatic ETL pipeline tool, we scraped, filtered, and transformed the data so that WISMIR3 differs from popular datasets like COCO and Flickr30k. We prove this difference by comparing linguistic statistics between the three datasets also computed using the tool. The purpose of WISMIR3 is to use it as a hard benchmark to challenge state-of-the-art text-image retrieval approaches, which already reach 90% Recall@5 scores on the mentioned popular datasets. With the experiments in this paper, we show that the text-image retrieval performance of the current models on WISMIR3 is much lower than on COCO or Flickr30k, as anticipated.

7 License

The dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) [6]. This allows copying and redistributing the data in any medium or format when appropriate credit is given and a link to the license is given. Further, it is allowed to mix, transform, or extend the dataset for any purpose. However, every change has to be indicated.

References

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, U.S.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*, pages 104–120, Online.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

James N. Farr, James J. Jenkins, and Donald G. Paterson. 1951. Simplification of Flesch Reading Ease Formula. *Journal of applied psychology*, 35(5):333.

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*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

[5] https://huggingface.co/openai/clip-vit-base-patch16

[6] https://creativecommons.org/licenses/by-sa/4.0/

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

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 *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajivc, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *arXiv preprint arXiv:2004.10643*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 1143–1151, Granada, Spain.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pages 140–153, Boston, MA, USA.

Florian Schneider, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards Multi-Modal Text-Image Retrieval to Improve Human Reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (NAACL SRW)*, Mexico City, Mexico (Online).

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, Melbourne, Australia.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Mike 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)*, Online.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2):64–73.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, Vancouver, Canada.

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 (TACL)*, 2:67–78.

Zhou Yu, Jing Li, Tongan Luo, and Jun Yu. 2020. A PyTorch Implementation of Bottom-Up-Attention. https://github.com/MILVLG/bottom-up-attention.pytorch.