

NewsImages: Addressing the Depiction Gap with an Online News Dataset for Text-Image Rematching

Andreas Lommatzsch
andreas.lommatzsch@dai-labor.de
DAI-Labor, TU-Berlin
Berlin, Germany

Benjamin Kille
Özlem Özgöbek
benjamin.u.kille@ntnu.no
ozlem.ozgobek@ntnu.no
NTNU
Trondheim, Norway

Yuxiao Zhou
Jelena Tešić
y_z37@txstate.edu
jtesic@txstate.edu
Texas State University
San Marcos, TX, USA

Cláudio Bartolomeu
David Semedo
c.bartolomeu@campus.fct.unl.pt
df.semedo@fct.unl.pt
Universidade NOVA de Lisboa
Lisbon, Portugal

Lidia Pivovarov
lidia.pivovarova@helsinki.fi
University of Helsinki
Helsinki, Finland

Mingliang Liang
Martha Larson
mliang@cs.ru.nl
mlarson@cs.ru.nl
Radboud University
Nijmegen, Netherlands

ABSTRACT

We present NewsImages, a dataset of online news items, and the related NewsImages rematching task. The goal of NewsImages is to provide researchers with a means of studying the *depiction gap*, which we define to be the difference between what an image literally depicts and the way in which it is connected to the text that it accompanies. Online news is a domain in which the image-text connection is known to be indirect: The news article does not describe what is literally depicted in the image. We validate NewsImages with experiments that show the dataset’s and the task’s use for studying occurring connections between image and text, as well as addressing the depiction gap, which include sparse data, diversity of content, and importance of background knowledge.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**;
• **Applied computing** → **Document management and text processing**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

datasets, multi-modal matching, depiction gap, embeddings

ACM Reference Format:

Andreas Lommatzsch, Benjamin Kille, Özlem Özgöbek, Yuxiao Zhou, Jelena Tešić, Cláudio Bartolomeu, David Semedo, Lidia Pivovarov, Mingliang Liang, and Martha Larson. 2022. NewsImages: Addressing the Depiction Gap with an Online News Dataset for Text-Image Rematching. In *13th ACM Multimedia Systems Conference (MMSys ’22)*, June 14–17, 2022, Athlone, Ireland. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3524273.3532891>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MMSys ’22, 14th–17th June, 2022, Athlone, Ireland

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/10.1145/3524273.3532891>

1 INTRODUCTION

Recent years have seen the expansion of multimedia research from image labeling [3, 11] to captioning [7] and cross-modal retrieval [23]. At each stage, researchers strive to expand our ability to model meaning across modalities. In order to continue to advance the state of the art, the research community requires datasets that represent the different ways in which images and text can be connected in natural settings. For example, in image labeling and image captioning, the emphasis is on text that directly captures the visually depicted content of the image. Also, many cross-modal retrieval tasks that the research community addresses, assume that the relationship between the textual query and the image is based on the explicit content of the image, i.e., illustration or exemplification. We introduce the *NewsImage* dataset and rematching task to push forward the state of the art with regard to the *depiction gap*—the difference between what an image literally depicts and the reason why it is connected to a text that it accompanies.

We choose online news because it is an important domain in which indirect connections between images and text arise naturally. Previous work [14] has provided evidence that the images accompanying online news articles should not be assumed to depict events described in the text. Specifically, [14] describes a case study on online news related to flooding in Southeast Asia and documents the displacement between the events described in a news item’s text and depicted in an image. For example, an article about a cancelled hydropower plant is accompanied by an image of workers installing electrical towers, which was taken months earlier to the events described in the article.

It is important to note that the depiction gap is part of the intentional process of creating an informative and engaging news story. An image that is complementary to what is described in the text contributes more to a news item than an image that conveys redundant information. The images accompanying news items are important for catching the readers’ attention, helping to visualize events happening, and putting the focus on certain aspects of a topic or people. Diverse types of images accompany news texts. Stock photos, portraits of people, photos from the place of events,

recently taken photos, photos belonging to an earlier related event exemplify this diversity. Editorial processes selecting accompanying images may differ among publishers. Some publisher may have an automated process in place to select images from a database with stock photos. Other publishers may task employees to assign images manually. Investigating the natural variation associated with the depiction gap will further multimedia research supporting investigation of the nature of indirect image/text connections. However, the ability to handle the depiction gap is also important for specific applications, such as systems that automatically select images to accompany articles, verify the trustworthiness of news, or leverage multimodal features to provide users with information access, via, e.g., search engines and recommender systems.

In this paper, we introduce the *NewsImages* dataset¹ and define the NewsImages rematching task, which support the study of the depiction gap in the online news domain. Fig. 1 highlights the indirect relationship between texts and images characteristic of the dataset. To address the NewsImages rematching task, systems must successfully reconstruct which images were originally published with which article. The dataset contains nearly 10 000 pairs of images and texts collected from a German newspaper, divided into a training and test set. Each pair consists of the title of a news article, the article's first lines (at most 256 characters), as well as an image that was published to accompany the article. In the test set, the images have been separated from the text to support the task of *Text-Image Rematching*. The NewsImages dataset has been validated at the 2021 edition of the Multimedia Evaluation Benchmark² within the *News Images Rematching* [8] task. The data set that we release contains both the split used for the benchmark and the ground truth. Thus, researchers can reproduce existing results as well as explore additional evaluation schemes.

2 RELATED WORK

In this section, we present a selection of related data sets for image and news analysis. The analysis shows that existing image datasets focus on textual descriptions of images while the datasets related to news focus on recommendations. In short, the examples we cover in this section affirm that existing datasets fail to provide adequate opportunities for studying the depiction gap.

2.1 Relating Text and Images

Datasets play an important role in the field of Artificial Intelligence (AI) and Machine Learning. Releasing new datasets challenges researchers to tackling new problems and improve the performance of existing AI algorithms. For instance, multimodal datasets lead to research on scene understanding. Cross-modal datasets have helped to explore the direct relationship between texts and images focusing on detecting visual concepts. In news, the relationship is rather indirect.

MS COCO (Microsoft Common Objects in Context) is a large-scale dataset for many tasks including object detection, segmentation, image captioning and cross-modal retrieval [11]. It contains

over 300 000 images. Five sentences describing the image's scene complement bounding boxes and per-instance segmentation masks.

The *Flickr30k* dataset contains over 30 000 images, each with five sentences given by annotators [23]. The dataset has been collected for a purely linguistic task; recently, it has been also used in cross-modal retrieval task. In contrast to our NewsImages dataset, both the Flickr30k and the MS COCO datasets describe images directly; whereas the NewsImages dataset focuses on the images embedded in news articles without considering image labels.

The *Visual Genome* dataset [9] provides images and complex labels. The image annotations provide region descriptions, objects, attributes, and relationships. Rich labels enable a variety of tasks, such as cross-modal retrieval and VQA (Visual question answering).

2.2 Multimodal News Datasets

To offer an alternative perspective on image description, the *GoodNews* dataset provides images with captions, headlines, and text articles collected using the New York Times API [2]. The dataset provides the URLs of the articles published in the New York Times from 2010 to 2018. It has 466 000 images, each with a caption. The *NYTimes800k* dataset [22] has a similar structure. The dataset provides longer and more complete captions resolving the missing text issue in the GoodNews dataset; it contains about 800 000 images with captions from The New York Times. Both GoodNews and NYTimes800k have been used for the task of image captioning.

2.3 News Datasets

The proposed NewsImages dataset has several connections to the NewsREEL dataset [12] originally published in the CLEF NewsREEL challenge. The NewsREEL dataset provides snippets from several German news portals and discussion groups. In addition to news snippets, information about the number of pages impressions, recommendations, and user clicks are provided. The data in NewsREEL is given in a stream (based on timestamps). The dataset has been optimized for developing news recommendation algorithms and method for predicting user preferences based on a data stream.

With NewsImages, the focus is on understanding the connection between news text and images. Thus, in contrast to the NewsREEL dataset, the NewsImages dataset assures user privacy, meaning that user access information such as click-data is excluded. As the NewsREEL scenario focused on trend prediction based on user-item interactions, the images are not included in the NewsREEL dataset.

Our NewsImages dataset supports the investigation of automatic concept detection and extraction from images and text. However, the NewsImages dataset and the rematching task go beyond existing datasets to introduce new challenges, such as handling sparsity and the loose semantic connection between images and texts. These challenges will be discussed further in Section 5. We also point out that the NewsImages dataset contains high quality regional news in the German language. As such, its content complements datasets of articles from much English-language newspapers with a much larger readership extending beyond their home city, like the New York Times.

¹The data set is available on Github: <https://github.com/NewsImagesDataset/NewsImagesDataset>

²<https://multimediaeval.github.io/editions/2021/>




(a)		In the face of what is possibly the worst tropical storm in decades, tens of thousands of people in southern Thailand have left their homes and sought protection. The residents on the coast of the province of Nakorn Si Thammarat, in the storm 'Pabuk' on Thursday.
(b)		The Bonn police speak of a "busy New Year's Eve", while the fire brigade and rescue service look back on a comparatively quiet New Year's Eve. The police operations control center counted 277 emergency calls and reports from 8 pm to 6 am.
(c)		Actually, Noel Diaz wanted to celebrate Christmas and New Year's Eve with his family. The young man from Aachen and his father José had planned to fly to Alicante (Spain) to visit relatives there. For Noel, traveling is not...

Figure 1: Example news item in the collection. These examples illustrate a typical case in which few concepts are shared between text and image. The texts are English translations of the provided German texts.

3 NEWSIMAGES DATASET

In this section, we describe the data set, including the design and the creation.

Requirements for the Dataset. The NewsImages dataset is designed to fulfill several requirements. We wanted it to be built based on real-world data collected specifically in the domain of online news. Further, a unique mapping between news articles and images should exist. This simplifies the evaluation, helping to avoid the issue of multiple possible ground truths encountered by ImageNet [3], while also maintaining the natural link between image and text.

Dataset Creation. The dataset has been created by collecting news article snippets and the links of the embedded images from major regional news portals (i.e., ksta.de) in Germany, which provide news about politics, sports, business, and local events. The dataset comprises news items published between January 2019 and April 2019 (4 months). Each dataset entry includes the URL, the item's title, and a text snippet of at most 256 characters as well as the URL to the image. The images have to be downloaded separately due to legal restrictions.

Dataset Annotation. The ground truth for our NewsImages Rematching task is straightforward to create. The relationship between the news article is extracted from the web portal. We merely separate the image from the original text in a large test set. Thus, the publisher defined the link between text and image. This simple approach ensures that we study the depiction gap as it naturally occurs. Such naturalness would be impossible, if human annotators were creating text or verifying specifically for the purpose of a multimedia task, as happens with other widely used data sets.

In order to create a valid dataset, we checked the elements of the dataset. We ensure that for all dataset news items, the title, the snippet, and the URL are present. In addition, we removed (near) duplicates ensuring that a 1:1 relationship between news articles and images exist. In other words, each image and each article is used exactly once.

Compared with the MediaEval 2021 version of this data set, which required a licensing agreement [8], the NewsImages dataset does not include click through data or access statistics as these data are not needed for the text-image rematching task. The item access

statistics had been used for another task concerning popularity prediction and recommendation. The data are provided as UTF8-encoded text files formatted as tab-separated data (TSV).

Dataset Statistics. The data set comprises four batches. The first three batches constitute the training data; the fourth batch the test data. The training data contains the links between articles and images. In the fourth batch (the test batch for Image-Text ReMatching), articles and images have been disassociated, which also indicates that the batch should be used for testing. Elsewhere, the release contains the information about the correspondence between articles and images in the fourth batch (i.e., the ground truth) to enable researchers to carry out independent evaluation of their approaches to the rematching task.

Table 1: Data Set Statistics. The data set is split into four batches. The number of cases refers to both articles and images (1-to-1 relation between news articles and images)

Batch	Purpose	Period	size
1	Training	Jan 2019	2539
2	Training	Feb 2019	2604
3	Training	Mar 2019	2387
4	Test	Apr 2019	1915

4 NEWSIMAGES REMATCHING TASK

The NewsImages rematching task is formulated as follows:

By using the news articles and accompanying images in the provided dataset, researchers should develop models that predict which image was published with a given news article.³

Text-image rematching requires addressing the depiction gap: a deep understanding of the news texts and the semantics of images is necessary. The rematching task differs from the tasks of labeling images and cross-modal retrieval. Generating image labels typically tries to identify entities and objects in images for getting a textual description of the images. News article-image retrieval tries to extract the central message of the news and to find semantically

³<https://multimediaeval.github.io/editions/2021/tasks/newsimages/>

relevant images: issues important for the depiction gap, such as image-text complementarity, are not addressed.

4.1 Metrics and Evaluation

To evaluate the effectiveness of rematching methods, the NewsImages rematching task defines two official metrics for quantifying the quality of text-image rematching. Recall, that we can assume that there is a 1:1 relationship between news articles and images. We conceptualize the rematching task as a ranking task, i.e., predicting a (sorted) list of images. The best matching images should be on the top of the list. For the metrics we adopt Mean Reciprocal Rank (MRR) and Mean Recall at N . MRR gives the best score, if the correct image is the top element in the prediction list; MRRN gives the highest score if the correct image is among the top N predictions. The metric Mean Recall at N reflects that several images may match a given news article. The case that a breaking news event temporarily lacks images motivates publishers to use stock photos. Stock photo databases can contain a plethora of images for some types of events. Focusing only on the top-ranked images would hide interesting models that rank equivalent images higher. Thus, we consider a set of cut-off points for the ranking metrics.

For completeness we include a formal specification of these metrics. Given a set of D news items and for each news text $d \in D$ a ranked list of images sorted based on the probability to match the text d , the rank_i describes the rank of the correct image for a given news text d_i . The metrics MRR and Mean Recall are computed as:

$$(1) \quad \text{MRR} = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{\text{rank}_i}$$

$$(2) \quad \text{Mean Recall at } N = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{1}_{\{\text{rank}_i \leq N\}}$$

The dataset release includes an evaluation component, computing the metrics MRR and Mean Recall at N . The component has been tested in the MediaEval benchmark 2021.

4.2 Challenges of NewsImages Rematching

The NewsImages rematching task, carried out on the NewsImages dataset poses several interesting research challenges related to the depiction gap. A key challenge is that strategies used by news portals to select images differ. For instance, whenever there are no photos immediately available for breaking news, publishers rely on stock photos or their photo archive, which could contain images from (more or less) related events.

The diversity of images used for visualizing news events constitutes another challenge. Some news items exhibit larger or lesser depiction gaps depending on their context. For example, a news message about a soccer game could have the portrait of a famous player, a scene of the game, a stock photo from the stadium, or the press conference. Understanding the relationship between text and image requires going beyond superficial knowledge of image content and text topic. In the next sections, we look at baselines for the rematching tasks and strategies for addressing these challenges.

5 DATASET VALIDATION AND APPLICATION

This section validates the NewsImages rematching task with baselines and exploratory approaches. The findings demonstrate the utility of the task for addressing the depiction gap. Table 2 provides

a summary and evaluation scores according to the task’s official metrics.

5.1 Baselines

We report two interesting baselines that were presented in short working notes papers at MediaEval 2021. The first baseline, described in [10], uses transfer learning, and demonstrates the contribution that model pre-training makes to the task. Semantic embeddings and pre-trained models can enrich the data and improve the matching precision, addressing sparsity in the dataset. The approach is based on ViT, which maps image patches to embeddings to create visual embeddings comparable to word embeddings [4]. Four datasets are used for pre-training: Microsoft COCO (MSCOCO), Visual Genome (VG), SBU Captions (SBU), and Google Conceptual Captions (GCC) [9, 11, 15, 19]. Comparing the first two lines of Table 2, labeled ‘Baseline using pre-training’, we see that the results confirm our expectation that pre-training delivers a clear advantage. The second baseline, implemented in [16], uses CLIP [17], which produces text and image embeddings in the same space. The third line of Table 2 shows that CLIP outperforms the transformer of [10].

5.2 Exploring the Depiction Gap

This section aims to provide insights into the extent to which the NewsImages dataset indeed involves a depiction gap. We look at the performance achieved by two approaches that attempt to leverage an explicit match between what is mentioned in the text and what is depicted in the images. The performance limitations of these approaches reveal the extent to which the NewsImages dataset pushes researchers to go beyond explicit mentions in text and literal depictions in images to address the depiction gap.

Direct-depiction matching. Four methods for calculating a match between images and texts were proposed in [25]. The methods leverage different correlations between text and image based on depicted content, and we refer to them as *direct-depiction matching methods*. They allow us to better understand the extent to which the NewsImages dataset includes pairs of text and images that are directly and indirectly related. Each approach is described in turn, following the overview in Fig. 2.

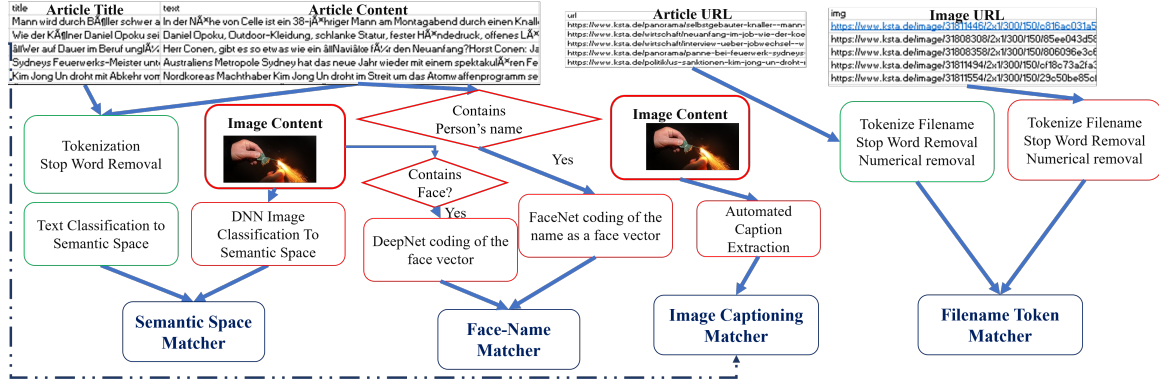
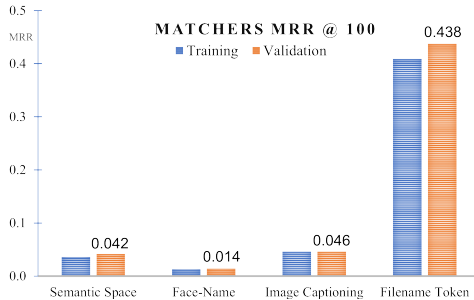
The *Semantic Space Matcher* uses a text classifier and an image classifier that both predict the same set of 70 categories, facilitating mapping images and text into compatible vectors that can be compared with cosine distance.

The *Face-Name Matcher* correlates the names within articles with the faces depicted in images using a 128-dimensional image-space embedding [24]. The Stanford Named Entity Recognizer (NER) [5] provides a Named Entity Recognizer for the person name extraction. FaceNet encodes the person’s name from the article as a 128-dimensional face vector [18], and DeepFace [20] encodes the detected faces in the image to the same space. The image is matched to the article based on a minimum cosine distance between the vectors for 24% of the articles that contain person names. For articles that lack person names, image captioning is used.

The *Image Captioning Matcher* adopts the captioning model pretrained with COCO dataset [13] for image caption generation. Word Mover’s Distance (WMD) is used to compare the similarity between image captions and article title.

Table 2: Baseline performance for the rematching task on the NewsImages dataset as well as the performance of approaches that explore or address the depiction gap between images and text.

Approach	Description	R@5	R@10	R@50	R@100	MRR@100
Baseline:						
Transformer [10]	no pre-training	0.017	0.039	0.126	0.218	0.017
	pre-training	0.079	0.134	0.334	0.465	0.059
Baseline:						
CLIP [16]	CLIP	0.220	0.300	0.530	0.640	0.169
Direct-depiction matching [25]	fusion method with URL	0.371	0.409	0.466	0.493	0.287
Context-enriched Transformers [1]	no face features nor named entities	0.127	0.190	0.438	0.587	0.093
	with face features and named entities	0.126	0.205	0.461	0.605	0.093
	fusion method	0.146	0.218	0.478	0.627	0.104
Visual Topic Models [16]	text teacher 120 topics	0.050	0.090	0.300	0.430	0.042
	joint 120 topics	0.060	0.090	0.260	0.400	0.043

**Figure 2: Data processing flow for four text-image matchers that assume different kinds of direct connections between what is depicted in the images and what is described in the text [25]****Figure 3: Comparison of the four direct-depiction matching methods on the training and validation datasets [25]**

The *Filename Token Matcher* compares the list of tokens extracted from the filename in the last part of text article URL with the lists of tokens extracted from the image filename in the last part of URL. The match calculates the number of overlapping tokens.

Fig. 3 illustrates the results of all four matchers on both the training and validation set. It can be seen that explicit matching

of the depicted content allows a certain level of success, but that these relatively simple approaches perform subpar to the approach of filename-token matching. In Zhou et al. [25], a fusion approach (combining all four approaches explained above) is reported as 'Direct-depiction matching'.

The filename-token matching leverages the filenames in the URL. The motivation for this approach is that human-created filenames contain human assigned semantic descriptions. They thus constitute an extra source of metadata, which narrows the match. This condition is interesting because it provides us an upper bound of what automatic approaches might achieve, if there was filename-related style information in the image text and images that was being leveraged. This point is important to keep in mind for future work.

Context-enriched Transformers. Bartolomeu et al. [1] proposed a set of context-enriched transformers to address the text-image rematching task. Like direct-depiction matching, this approach attempts to leverage correlations between images and text related to explicitly mentioned and literally depicted content. Here, the architecture is more sophisticated. The context-enriched transformers are variants of a multimodal transformer model, based

on LXMERT [21]. The idea is to provide complementary views of the two modalities, and leverage the model’s capability of jointly attending to different news elements when predicting. In practice, a multimodal context-enriched sequence is created, comprising image regions and faces—from the visual modality—and news text and named entities—from the textual modality. This sequence is then given as input to an LXMERT-based model, adapted to learn individual projections for each type of sequence context elements.

In Bartolomeu et al. [1], it is assumed that named entities play a major role in defining the news topic and scope. These entities are taken to define context, and the context-enriched transformer approach seeks to exploit such named entities and their visual depictions in the images. First, the approach increases the importance of all entity tokens. Second, [1] observed that journalists often chose images depicting people mentioned in the news items. Consequently, they augmented the model’s inputs to incorporate both face features and named entities, such that the model is allowed to relate entities and faces. Table 2, ‘Content-enriched transformers’ reports the effectiveness of the approach with and without the two elements (faces and entities). For higher K , accounting for entities and faces is beneficial, while $R@5$ equivalent performance is observed. These results demonstrate that while explicitly matching named entities to literally depicted image content addresses some of the news-image connection scenarios, the NewsImages dataset indeed involves data with a large, and diverse depiction gap, thus requiring more expressive models.

5.3 Addressing the Depiction Gap

In this section, we analyze the Visual Topic Model, an approach developed by Pivovarov and Zosa [16] to address the Image-Text Rematching task. We discuss this approach in order to illustrate the potential of the dataset for inspiring creative ideas of how to handle the varied and indirect nature of naturally occurring image-text connections. Note that we do not consider these approaches to have solved the problem. Rather, it provides evidence of directions that can be productively explored using the dataset.

The Visual Topic Model (VTM) is trained on pairs of aligned text and image. During inference, it is able to generate a topic distribution for an image, without using any text. Each topic constitutes a distribution over the vocabulary. For instance, Fig. 4 shows an example image with the top words associated with the most probable topic. As could be seen from the words, the topic is associated with football, thus an image is labeled with football theme and VTM could work as an image-labelling tool.

For the text-image rematching task, image and text topics should be aligned. To obtain aligned visual and textual topics, Pivovarov and Zosa [16] used a knowledge distillation. In this approach, first a textual teacher model is trained; then an image model is trained as a student. A pre-trained teacher takes as an input text and outputs its topic distribution. A student takes the corresponding image and produces a topic distribution close to the textual one. Alternatively, one could first train a joint model, which takes as input both text and image, and then train two student models, one for text and one for image. Table 2 reports the results as ‘Visual Topic Model.’

Since the joint model uses all available information, it could, in theory, describe the data better than the text-only one. Therefore,



trainer, leverkusen, bayern, league, haie, fortuna, bundesliga, sieg, spiel
trainer, leverkusen, bavaria, league, sharks, fortune, bundesliga, victory, game

Figure 4: An example image and top words associated with the most probable topic for that image.

it is expected that it outperforms the text teacher concerning $R@5$. However, it is interesting to note that the text-teacher performs superior with respect to $R@50$ and $R@100$. This difference should be investigated in future work, to determine the extent to which the use of images during training adds noise. The VTM does not outperform the baseline, but the example in Fig. 4 suggests that it is able to represent images going beyond the visually depicted content, which means it remains an interesting approach to explore in the future. VTM could be useful for other tasks, for example, to illustrate to an author diverse images for their article, i.e. sampled from different topics.

6 CONCLUSION AND OUTLOOK

We have introduced the NewsImages dataset and proposed an associated image-text rematching task. We have argued that the dataset will encourage researchers to tackle the task of image-text rematching, and that this task will help them to come to grips with the depiction gap, i.e., the difference between what an image literally depicts and the way that it is connected to a text that it accompanies.

Looking forward to the future, an advantage of the NewsImages dataset is the ease with which it can be expanded. For example, the dataset can be extended by integrating additional news portals, such as those in e.g., the Adressa Dataset [6]. As mentioned, it is important to make sure that there is a 1:1 correspondence between images and text in order to be able cleanly interpret results. Also, it would be easy to add more text beyond the first 256 characters of the news item to the dataset. This extension would support investigation how an image relates to the large topic of a news item. Crucially, extending the data set does not involve manual annotation because of our simple but useful procedure of creating ground truth by disassociating text and image.

An improved understanding of the whole range of relationships that can connect images and text is important for multimedia research. Practically, such an understanding could improve systems that identify misinformation by identifying images that do not truthfully match the text that they accompany. Further, better understanding of the relationships that can exist between image and text will help us build explainable multimodal systems, specifically cross-modal retrieval systems or multi-modal recommender systems.

REFERENCES

- [1] Cláudio Bartolomeu, Rui Nóbrega, and David Semedo. 2021. NewsSeek-NOVA at MediaEval 2021: Context-enriched Multimodal Transformers For News Images Re-matching. In *Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021* (Online). CEUR Workshop Proceedings.
- [2] Ali Furkan Biten, Lluís Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12466–12475. <https://doi.org/10.1109/CVPR.2019.01275>
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNET: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua 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 *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2010.11929>
- [5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (Ann Arbor, Michigan) (ACL '05). Association for Computational Linguistics, USA, 363–370. <https://doi.org/10.3115/1219840.1219885>
- [6] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The Adressa Dataset for News Recommendation. In *Proceedings of the International Conference on Web Intelligence*. 1042–1048. <https://doi.org/10.1145/3106426.3109436>
- [7] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 417–434. https://doi.org/10.1007/978-3-030-58520-4_25
- [8] Benjamin Kille, Andreas Lommatzsch, Özlem Özgöbek, Mehdi Elahi, and Duc-Tien Dang-Nguyen. 2021. News Images in MediaEval 2021. In *Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021* (Online). CEUR Workshop Proceedings.
- [9] 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* 123, 1 (2017), 32–73. <https://doi.org/10.1007/s11263-016-0981-7>
- [10] Mingliang Liang and Martha Larson. 2021. Exploring a Pre-trained Model for Re-Matching News Texts and Images. In *Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021* (Online). CEUR Workshop Proceedings.
- [11] 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*. Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
- [12] Andreas Lommatzsch, Benjamin Kille, Frank Hopfgartner, Martha Larson, Torben Brodt Özlem Özgöbek, and Jonas Seiler. 2017. CLEF 2017 NewsREEL Overview: A Stream-based Recommender Task for Evaluation and Education. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction; Proceedings of the 8th International Conference of the CLEF Association, CLEF 2017* (Dublin, Ireland) (LNCS, vol. 10456). Springer International Publishing, 239–254. <https://doi.org/10.1007/978-3-319-65813-1>
- [13] Ruotian Luo, Gregory Shakhnarovich, Scott Cohen, and Brian Price. 2018. Discriminability Objective for Training Descriptive Captions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6964–6974. <https://doi.org/10.1109/CVPR.2018.00728>
- [14] Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. 2020. The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4343–4351. <https://aclanthology.org/2020.lrec-1.535>
- [15] Vicente Ordóñez, Girish Kulkarni, and Tamara L Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* (Granada, Spain) (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 1143–1151. <https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7e9ea9-Paper.pdf>
- [16] Lidia Pivovarov and Elaine Zosa. 2021. Visual Topic Modelling for NewsImage Task at MediaEval 2021. In *Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021* (Online). CEUR Workshop Proceedings.
- [17] 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. In *Proceedings of the 38th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823. <https://doi.org/10.1109/CVPR.2015.7298682>
- [19] 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, 2556–2565. <https://doi.org/10.18653/v1/P18-1238>
- [20] Yaniv Taigman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1701–1708. <https://doi.org/10.1109/CVPR.2014.220>
- [21] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 5100–5111. <https://doi.org/10.18653/v1/d19-1514> arXiv:1908.07490
- [22] Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13035–13045. <https://doi.org/10.1109/CVPR42600.2020.01305>
- [23] 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. https://doi.org/10.1162/tacl_a_00166
- [24] Zhao Yumeng, Yun Jing, Gao Shuo, and Liu Limin. 2021. News Image-Text Matching With News Knowledge Graph. *IEEE Access* 9 (2021), 108017–108027. <https://doi.org/10.1109/ACCESS.2021.3093650>
- [25] Yuxiao Zhou, Andres Gonzalez, Parisa Tabassum, and Jelena Tesic. 2021. DL-TXST NewsImages: Contextual Feature Enrichment for Image-Text Re-matching. In *Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021* (Online). CEUR Workshop Proceedings.