FEIDEGGER: A Multi-modal Corpus of Fashion Images and Descriptions in German

Leonidas Lefakis, Alan Akbik, Roland Vollgraf
Zalando Research
Mühlenstraße 25, 10243 Berlin
{firstname.lastname}@zalando.de

Abstract

The availability of multi-modal datasets that pair images and textual descriptions of their content has been a crucial driver in progress of various text-image tasks such as automatic captioning and text-to-image retrieval. In this paper, we present FEIDEGGER, a new multi-modal corpus that focuses specifically on the domain of fashion items and their visual descriptions in German. We argue that such narrow-domain multi-modality presents a unique set of challenges such as fine-grained image distinctions and domain-specific language, and release this dataset to the research community to enable study of these challenges. This paper illustrates our crowdsourcing strategy to acquire the textual descriptions, gives an overview over the FEIDEGGER dataset, and discusses possible use cases.

Keywords: Crowdsourcing, Multi-modalily, Fashion, German

1. Introduction

Recent years have seen a renewed interest in text-image multi-modality and have seen the emergence of tasks such as automatically generating textual captions for a given image (Karpathy and Fei-Fei, 2015; Lu et al., 2016), using plain text descriptions to query images (Socher et al., 2014), and matching lexical tokens or constituents to regions in an image (Ma et al., 2015). This interest is driven by advances in multi-modal deep learning for computer vision (Vinyals et al., 2015; Karpathy et al., 2014) on the one hand, as well as the availability of paired text-image datasets on the other.

Multi-modal text-image datasets. Commonly cited multi-modal datasets either consist of user-captioned images from the web, such as FLICKR (Hodosh et al., 2013; Plummer et al., 2015) and online news (Hollink et al., 2016), or images for which crowd workers have produced visual descriptions (Rashtchian et al., 2010), such as the popular COCO dataset (Lin et al., 2014). In both cases, the textual data directly describes image content, thus enabling the above-mentioned lines of research.

However, these datasets are often restricted to English language text and typically of relatively broad domain; The FLICKR caption datasets for instance contain images including landscapes, animals, and everyday scenes while the COCO dataset is similarly broad but contains more items per image. This makes such datasets difficult to apply for study of multi-modality in more narrow domains. In the domain of fashion items for instance, images are broadly similar and are often distinguished only by fine-grained differences such as the material, the neckline, brand logos, the cut and the style of the hem. Similarly, the language used in fashion is domain-specific, tailored specifically to highlight such fine-grained differences. We argue that such narrow domains present unique research challenges that require specialized multi-modal datasets.

A multi-modal text-image corpus for fashion. With this paper, we introduce a novel dataset for research in narrow-domain multi-modality, called FEIDEGGER[1]. Contrary to previous datasets, we restrict the domain to images of one type of fashion item, namely dresses, and German-language visual descriptions. The dataset consists of 8,700 fashion items, each with a high resolution image and 5 independently collected textual descriptions of the item. The images are of each fashion item alone in front of a white background. Crowd workers were instructed to inspect each image and then produce a plain text description of the fashion item. For an example item in this dataset, see Figure 1.

In the remainder we provide a description of our dataset and describe the design of our crowdsourcing approach. We qualitatively analyze the obtained data and find that the crowdsourced descriptions are of high quality and are finely detailed, which we attribute to our careful choices of experimental parameters. Finally, we provide a description of how we package this data and discuss expected use cases. We release FEIDEGGER to the research community to further research in narrow-domain multi-modality.

2. Dataset Creation

In constructing FEIDEGGER we employed a crowdsourcing approach to produce accurate and succinct descriptions of each fashion image that make reference to fine-grained non-generic image features. Our pipeline required careful monitoring of individual workers' performance using automated evaluation of test-questions as well as performing pilot studies. However, we did not place very high demands on language correctness in terms of spelling and grammar. Rather we accepted average language use as might be expected in user reviews or forums on the web. The details of our pipeline are given in the following sections.

2.1. Task Design: Pilot Study

We first conducted a pilot study using the crowdsourcing platform CROWDFLOWER[2] to test the design of our crowdsourcing task and identify potential quality issues.

[1] A rough acronym of "fashion image data and descriptions in German"

[2] https://www.crowdflower.com/

Image 1

Description 1

Langes weißes Kleid mit Bugs Bunny Musterung an der Seite des Kleides. runder Ausschnitt und kurze Ärmel.

(engl.) Long white dress with Bugs Bunny pattern at the side of the dress. Round neckline and short sleeves.

Description 2

Schlauchkleid in weiß mit sehr kurzen Armen und einem Bugs Bunny Aufdruck auf der linken Seite. Arme und Hals haben einen schwarzen Streifen.

(engl.) Tube dress in white with very short sleeves and a Bug Bunny print on the left side. The sleeves and the neck have a black stripe.

Figure 1: Example item in FEIDEGGER: For each fashion item we provide an image and 5 crowdsourced descriptions (only 2 presented here). The image is always a still of the fashion item itself in front of a white background. Textual descriptions are in German and typically consist of 2-4 short sentences. English translations are provided for the purpose of illustration in this Figure, but are not part of the dataset.

Task design. Given the image of a fashion article workers were instructed to provide a German language description of the item. They were instructed to go into detail and write about 5 sentences. In order to discourage non-native speakers to participate in the task, we provided instructions in German language only and required workers to first complete a German-language tutorial.

Study parameters and results. We conducted the initial study over 1000 fashion items and restricted workers to a maximum of 50 descriptions each. We restricted the pool of workers to (a) those based in a German-speaking country and (b) *level 3* workers, who are the highest ranked workers according to the internal CROWDFLOWER system of evaluation.

Upon manual inspection of the results we found the produced textual descriptions to be of a very high quality, likely due to our very restrictive parameters in selecting workers. However, we also identified the following issues:

Short descriptions Although instructed to provide descriptions at reasonable length, some crowd workers provided very short descriptions, sometimes only a few words in length.

Unspecific descriptions We found some descriptions of reasonable length to be generic descriptions that some crowd workers simply re-used for each fashion item, sometimes directly copied from the tutorial examples.

Non-German text Finally, there were a number of instances in which workers had responded in another language than German, such as Polish.

While the issues of short and non-German crowd answers are fairly straightforward to address with automatic verification methods, the problem of catching workers that provide unspecific, non-matching or low-quality descriptions proved inherently more difficult. We therefore decided to adopt two strategies for increasing quality, namely curating workers and automatic quality checks. In the next two sections, we give an overview of each.

2.2. Building a Pool of Curated Workers

Our first and most important measure was to identify a pool of workers that could consistently and reliably create content to the level of quality we required. To this end, we constructed a crowdsourcing task that was open to any worker meeting the minimum requirements, hereafter referred to as the *trial task*. This task mirrored the task design of the actual crowdsourcing, but restricted each worker to provide a maximum of 100 descriptions, which we treated as a sample of the workers' ability to provide high quality descriptions. This sample was then manually assessed by experts.

Results. The task was executed for a period of 3 weeks. Approximately 150 distinct crowd workers participated in the task, of which we admitted 50 into the pool of curated workers. These workers were used for the final crowdsourcing task of generating image descriptions.

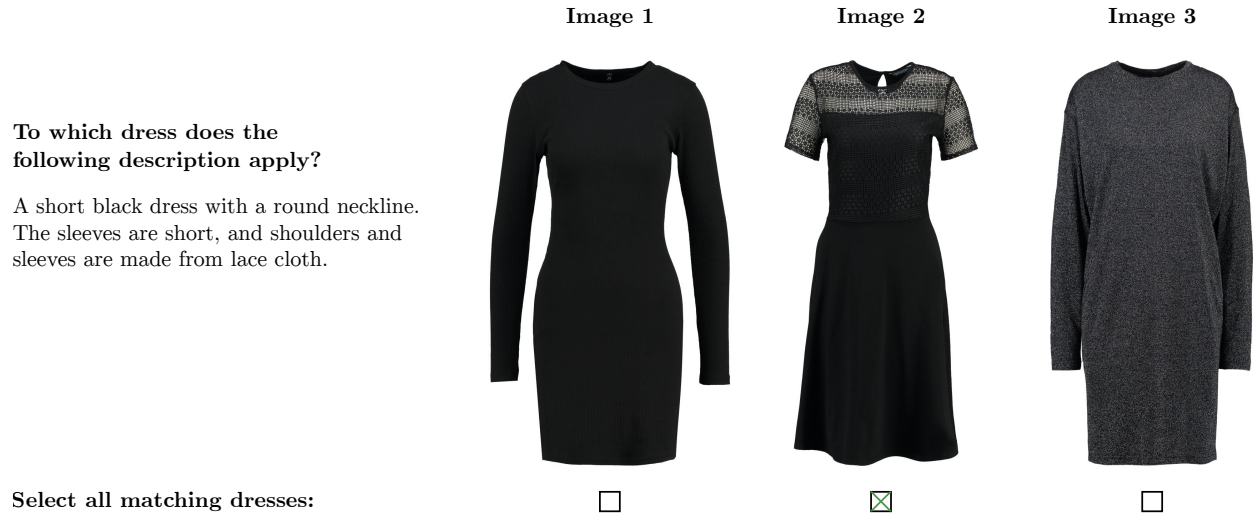

Figure 2: Example evaluation task. A description is paired with three images, one of which is the original source image for which the description was generated. The evaluation worker is instructed to select all images to which the description applies. In this example, the worker selects *image 2*, which is the correct source. *Note: Original task is in German, this example was translated for readability.*

2.3. Automatic Quality Control

In addition to curating workers, we employed three simple methods for automatic quality control to ensure that workers would not submit short or unrelated descriptions.

The first was to add a simple regular expression that checked whether each submitted text contained at least 10 distinct words. Similarly, the second checked whether at least one word of the submitted text was a German stopword, to identify responses that were either in a different language or not well-formed. Though these checks were rather coarse, they were nonetheless successful in filtering workers and descriptions to those of reliable quality. A manual inspection of both the cleared and rejected results showed that no non-German text was erroneously accepted while in only one case was work in German rejected as non-German.

The third was to introduce a series of test questions within the data presented to the workers. These consisted of images that had been manually inspected and for which a series of words at least one of which should appear in any reasonable description had been determined. For instance, for the images in Figure 1, we would expect the German words for black (*schwarz*) and white (*weiß*) to be mentioned in the description, as well as some synonym of either rabbit or Bugs Bunny. Workers that failed automatic quality control were removed from the pool and their work discarded.

2.4. Full Crowdsourcing Task

After building up the pool of curated workers and establishing the quality controls, we ran a large crowdsourcing experiment. Our goal was to annotate 8,764 fashion items with 5 descriptions each. Annotation was completed after 4 weeks of crowdsourcing.

3. Quality Estimation

In order to assess the quality of the gathered descriptions, we set up a separate crowdsourcing task. Since our main goal in creating FEIDEGGER was to acquire detailed, non-generic descriptions, the evaluation task was set up to evaluate both whether the descriptions were accurate *and* discriminative.

Evaluation task. To accomplish this, we designed the *evaluation task* as follows: Each crowd-generated description was paired with three images of fashion articles. One of the three images was the *source image*, i.e. the image for which the description had been produced. The other two images were other items that are visually similar to the source image, determined using pre-computed image embeddings over a large fashion catalogue (Bracher et al., 2016). One such example pairing of a description and three similar images is depicted in Figure 2.

Given such a pairing, a crowd worker was asked to select all images to which the description applies. Note that the worker was not informed that the description's origin was

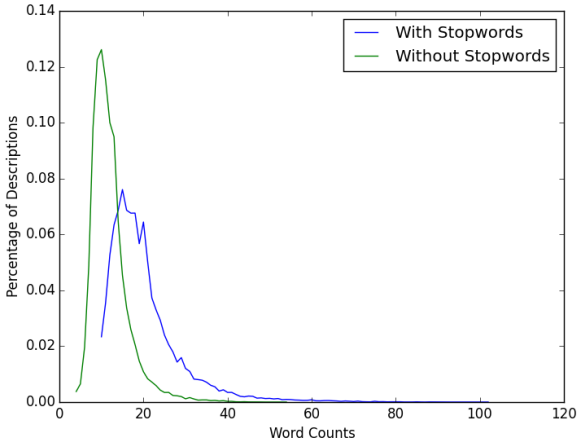

Figure 3: Plot of word count frequencies.

only of one of the images and was given the choice to select more than one image. Thus the task tested both for correctness (whether the source image was chosen) and discriminativeness (whether other images beyond the source image were also chosen). Figure 2 illustrates one such evaluation task.

Experimental results. We ran experiments on 4,000 description-image triplets. In 96.5% of cases the worker deemed the description relevant to at least one of the images. Furthermore, in 97% of those cases the worker picked the correct image as being relevant to the description, while in 96.35% of the cases the worker chose only that image. These results indicate that descriptions are generally of high quality and discriminatively match the source image.

Length of descriptions. To give an overview of the length of crowd-provided descriptions, we computed statistics on word count, as illustrated in Figure 3: Descriptions have an average total of 20.26 words, with a median of 18, and consist of, on average, 2.23 sentences. Stopwords make up roughly 40% of the data.

4. Data Release and Outlook

We release the data via a specifically dedicated website. This data may be useful for experiments various text-image tasks such as captioning and image retrieval, to compare the quality of approaches that work well on general datasets such as COCO and FLICKR with narrow-domain data, and to research approaches that work well in this domain.

Future work will focus on extending the scope of the crowdsourced image description along two dimensions. On the one hand, we will include other types of fashion items besides dresses, such as shoes and shirts. On the other hand, we aim to repeat crowdsourced data gathering efforts for languages other than German, such as English, French, and Dutch. The medium-term goal is to create a multi-modal dataset that for each fashion item contains descriptions in several languages.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 732328 ("FashionBrain").

5. Bibliographical References

Bracher, C., Heinz, S., and Vollgraf, R. (2016). Fashion dna: Merging content and sales data for recommendation and article mapping. *arXiv preprint arXiv:1609.02489*.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hollink, L., Bedjeti, A., van Harmelen, M., and Elliott, D. (2016). A corpus of images and text in online news. In *LREC 2016, 10th International Conference on Language Resources and Evaluation*.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Karpathy, A., Joulin, A., and Li, F. F. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2016). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*.

Ma, L., Lu, Z., Shang, L., and Li, H. (2015). Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2623–2631.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.