Horizon 2020

Understanding Europe's Fashion Data Universe

Enriched Image Dataset

Deliverable number: D6.4

Version 2.0

Funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 732328

Project Acronym:	FashionBrain
Project Full Title:	Understanding Europe's Fashion Data Universe
Call:	H2020-ICT-2016-1
Topic:	ICT-14-2016-2017, Big Data PPP: Cross-sectorial and cross-lingual data integration and experimentation
Project URL:	https://fashionbrain-project.eu

Deliverable type	Other (O)
Dissemination level	Confidential (C)
Contractual Delivery Date	31 December 2018
Resubmission Delivery Date	4 February 2019
Number of pages	17, the last one being no. 13
Authors	Alan Akbik, Roland Vollgraf - Zalando
Peer review	Alessandro Checco, Jennifer Dick - USFD

Change Log

Version	Date	Status	Partner	Remarks
0.1	19/11/2018	Draft	Zalando	
0.2	14/12/2018	Full Draft	Zalando	
1.0	20/12/2018	Final	Zalando, USFD	Rejected 30/01/2019
2.0	04/02/2019	Resubmitted Final	Zalando, USFD	

Deliverable Description

This deliverable consists of taking the data provided in D6.1 and enriching it with human labelling where data quality was insufficient.

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 deliverable, we present FEIDEGGER, a new multi-modal corpus that focuses specifically on the domain of fashion items and their visual descriptions in German. FEIDEGGER extends the dataset of dresses presented in Deliverable D6.1 by adding a layer of textual descriptions of visual features. Such text-image data is important to core FashionBrain use cases, as it creates a link between text and image features that can be used to train neural end-to-end text-to-image search approaches.

We argue that the narrow-domain multi-modality we investigated during the creation of FEIDEGGER presents a unique set of challenges such as fine-grained image distinctions and domain-specific language. We release this dataset to the project consortium (as well as the wider research community on request) to enable study of these challenges. Our approach has also led to a scientific publication of the paper "FEIDEGGER: A Multi-modal Corpus of Fashion Images and Descriptions in German" at the 11th Language Resources and Evaluation Conference (LREC), 2018. This deliverable illustrates our crowdsourcing strategy to acquire the textual descriptions, gives an overview over the FEIDEGGER dataset, and discusses possible use cases. We also describe dataset statistics in detail.

Table of Contents

List of Figures

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 [4, 7], using plain text descriptions to query images [11], and matching lexical tokens or constituents to regions in an image [8]. This interest is driven by advances in multi-modal deep learning for computer vision [12, 5] on the one hand, as well as the availability of paired text-image datasets on the other.

For the domain of Fashion and the goals of the FashionBrain project, such multi-modal data is crucial to train text-to-image search approaches. Unlike the dataset presented in Deliverable D6.1 which pairs images only with a fixed vocabulary of structured attributes (like color tags), multi-modal datasets employ the full breath of human language in textual descriptions.

Multi-modal text-image datasets. Commonly cited multi-modal datasets either consist of user-captioned images from the web, such as Flickr [2, 9] and online news [3], or images for which crowd workers have produced visual descriptions [10], such as the popular Coco dataset [6]. 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 deliverable, we introduce a novel dataset for research in narrow-domain multi-modality, which we call 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 adds a layer of annotations to the dataset of dresses released with Deliverable D6.1. It 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

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

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 1.1.

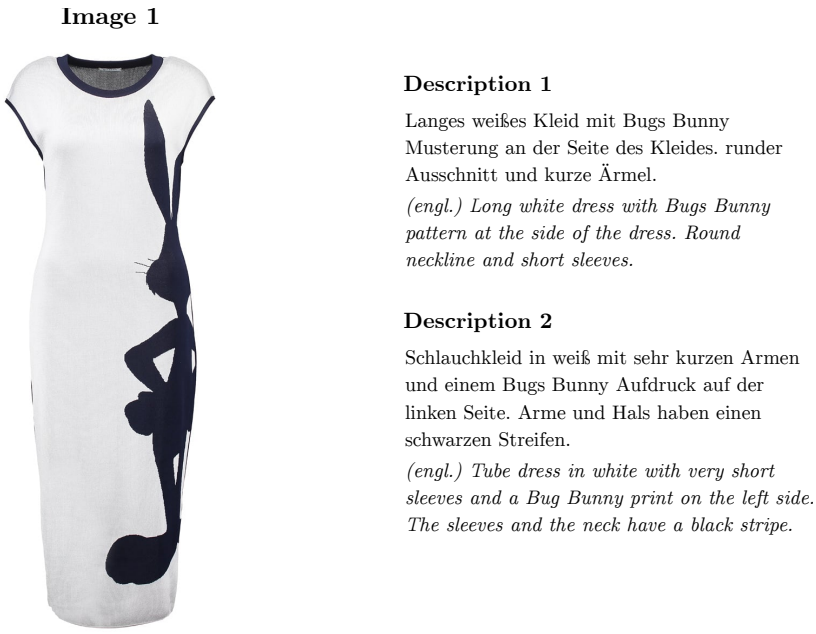

Figure 1.1: Example item in FEIDEGGER: For each fashion item we provide an image and 5 crowdsourced descriptions (only 2 presented here). The image is always a still of the fashion item itself in front of a white background. Textual descriptions are in German and typically consist of 2-4 short sentences. English translations are provided for the purpose of illustration in this Figure, but are not part of the dataset.

1.1 Placement of this Deliverable within FashionBrain

In this deliverable, *D6.4 "Enriched Image Dataset"*, we detail how we used crowdsourcing to enrich the dataset presented in D6.1 with textual descriptions for the purpose of research into text-image multimodality and, ultimately, text image search for deliverables D6.3 and D6.5.

As such, the requirements for the dataset are derived from the business scenarios identified in *D1.2 "Requirement analysis document"*, in particular:

- Scenario 1: End-To-End Multi-language Search
  - Challenge 1: Mapping Search Intentions to Product Attributes
  - Challenge 2: End-To-End Learning

This deliverable was created in a collaboration of project consortium partners. Partners USFD and UNIFR provided their expert know-how in crowdsourcing

to Zalando which prior to FashionBrain had little experience in setting up and executing crowdsourcing tasks. All experiments were designed in close collaboration and executed over the dataset provided in Deliverable D6.1.

1.2 Outline of this Deliverable Report

In this deliverable, we give details on:

- How we designed the crowdsourcing task, with help of crowdsourcing expertise provided by project consortium members (USFD, UNIFR). We detail the platform used to execute the experiments, and illustrate the scope.
- How we addressed the crucial question of assessing the quality of crowdsourced annotations, again leveraging expertise provided by project consortium members (USFD, UNIFR).
- The resulting dataset of this work, which we make available to the project consortium. Note that this deliverable is classified as confidential, meaning that the dataset cannot be shared with third parties without our express consent.

The remainder of this deliverable is as follows. We describe the design of our crowdsourcing approach, the execution of the crowdsourcing experiments and the results of a quality evaluation in Section 2. We then discuss the resulting dataset, illustrate use cases and dataset statistics in Section 3. Finally, we conclude this report and list next steps in Section 4.

2 Creation of the FEIDEGGER Dataset

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[1] to test the design of our crowdsourcing task and identify potential quality issues.

Task design. We presented an image of a fashion article to crowd workers and instructed them to provide a German language description of what they see in the image. 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.

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

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. The descriptions were manually assessed and each worker either cleared or rejected.

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.

2.3 Automatic Quality Control

Another result of the pilot study was to identify common quality issues in the crowdsourced data. To address these issues, we employed three simple methods for automatic quality control to ensure that workers would not submit short or unrelated descriptions. These measures were employed in addition to curating workers as discussed in the previous section:

Minimum length filter The first was to add a simple regular expression that checked whether each submitted text contained at least 10 distinct words. This simple filter addressed the problem of some crowdworkers submitting only very little text, presumably to complete tasks faster.

German language filter 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.

Test questions 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 1.1, 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 all 8,764 fashion items in the dresses dataset presented in Deliverable D6.1 with 5 distinct textual description each. We ran experiments in three batches and completed annotation after 1 month of crowdsourcing.

This yielded the FEIDEGGER dataset.

2.5 Quality Estimation

After completing crowdsourcing for FEIDEGGER, we needed to assess the quality of the gathered descriptions. To do this, 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 [1]. One such example pairing of a description and three similar images is depicted in Figure 2.1.

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 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 2.1 illustrates one such evaluation task.

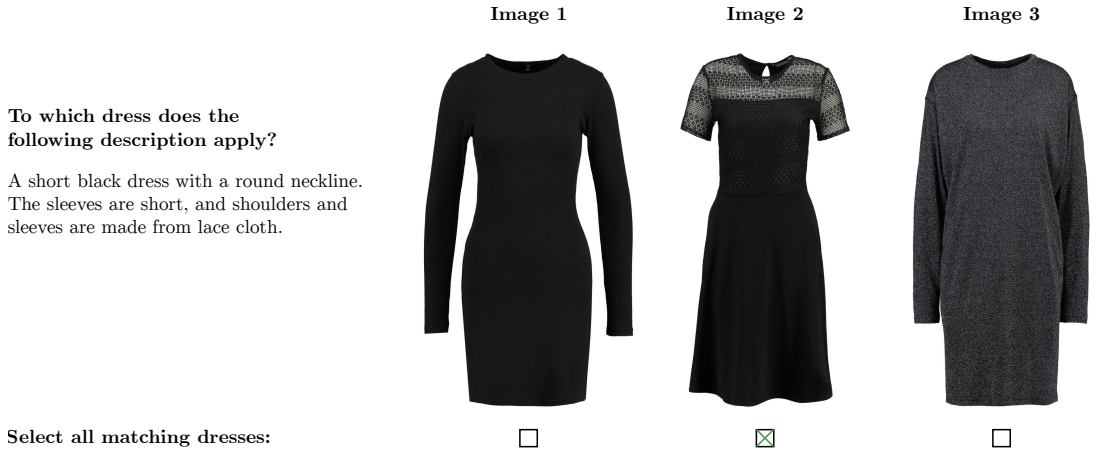

Figure 2.1: Example evaluation task. A description is paired with three images, one of which is the original source image for which the description was generated. The evaluation worker is instructed to select all images to which the description applies. In this example, the worker selects *image 2*, which is the correct source. *Note: Original task is in German, this example was translated for readability.*

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.

We note that due to the crowd-sourced nature of this evaluation, it is not necessarily true that in the 3% remaining cases the description was not relevant as this may be on error on the evaluating workers part. However in the 96.35% of the cases where the "target" image was marked as the only one relevant by the worker it is less likely that this was accidental on his part. As such we take this 96.35% to be a lower bound on the quality of the descriptions in the dataset.

These results indicate that descriptions are generally of high quality and discriminatively match the source image.

3 Dataset Release

We release the dataset to the project consortium and to the wider research community on request. For third party researchers, we believe 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. We announced this dataset to the research community via the following scientific publication:

- FEIDEGGER: A Multi-modal Corpus of Fashion Images and Descriptions in German. Leonidas Lefakis, Alan Akbik and Roland Vollgraf. *11th Language Resources and Evaluation Conference, LREC 2018.*

3.1 FashionBrain Use Cases

Within FashionBrain, this data will be used to train Zalando's end-to-end text-to-image search approach and so used to realize Scenario 1, "End-To-End Multi-language Search", as defined in the requirement analysis document Deliverable D1.2. It will be used to address the two challenges defined under this scenario, namely mapping search intentions to product attributes and end-to-end learning. We foresee that end-to-end learning in particular will benefit from this dataset, as it provides a different textual view on fashion items and thus enable research into more complex linguistic search queries for fashion items. These concepts will be further explored in Deliverable D6.5.

3.2 Dataset Statistics

FEIDEGGER consists of 43,840 textual descriptions. To give an overview of the length of crowd-provided descriptions, we computed statistics on word count, as illustrated in Figure 3.1: Descriptions have an average total of 20.26 words, with a median of 18, and consist of, on average, 2.23 sentences. Stop words make up roughly 40% of the data.

Next to textual descriptions, we also collect the answer to a simple yes/no question on whether the crowd worker considers the item in the image a 'dress'. This question was ansered with 'no' in only 8 instances throughout the entire dataset. Figure 3.2 provides an example of an item classified as 'dress' in the Zalando database that a crowd worker did not think was a dress.

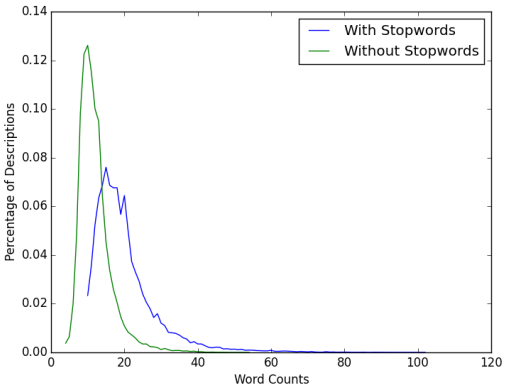

Figure 3.1: Plot of word count frequencies in textual descriptions.

Figure 3.2: Example item in FEIDEGGER: This item is classified as 'dress' in the Zalando database, but marked as 'not a dress' by one of the crowd workers, perhaps due to its bathrobe-like look.

3.3 Extending the Dataset

Future work (during and post-FashionBrain) 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. A medium-term goal is to create a multi-modal dataset that for each fashion item contains descriptions in several languages.

4 Conclusion and Outlook

This deliverable described the creation of a multi-modal text-image dataset specific to the domain of fashion. It adds a layer of textual descriptions to the dataset of dresses released in Deliverable D6.1: For each dress, we collected 5 independently-authored textual descriptions from crowd workers. Workers were instructed to describe what they see in the image and go into as much detail as possible. The crowdsourcing task was designed in collaboration between project consortium members Zalando, USFD and UNIFR and evaluated in a separate crowdsourcing task that showed annotations to be of high quality.

The resulting dataset, called FEIDEGGER, will enable research and development work into end-to-end search by providing labeled training data that links images to detailed, fulltext descriptions of visual features. This is of core significance to scenario 1 of the requirement analysis deliverable D1.2 and the overall goal of work package 6, namely "making images searchable by text". We expect that this dataset will enable more complex search queries in the final end-to-end search system that will be presented in Deliverable D6.5.

Bibliography

[1] Christian Bracher, Sebastian Heinz, and Roland Vollgraf. Fashion dna: Merging content and sales data for recommendation and article mapping. *arXiv preprint arXiv:1609.02489*, 2016.

[2] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[3] Laura Hollink, Adriatik Bedjeti, Martin van Harmelen, and Desmond Elliott. A corpus of images and text in online news. In *LREC 2016, 10th International Conference on Language Resources and Evaluation*, 2016. ISBN 978-2-9517408-9-1.

[4] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[5] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[7] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2016.

[8] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2623–2631, 2015.

[9] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 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, 2015.

[10] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 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, 2010.

[11] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[12] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.