

Intermediate Fusion for Multimodal Product Matching

Jacob Pollack^{1,*}, Hanna Köpcke² and Erhard Rahm¹

¹Leipzig University & ScaDS.AI, Leipzig, Germany

²University of Applied Sciences Mittweida, Mittweida, Germany

Abstract

Web-based entity resolution, particularly in the context of online marketplaces and e-commerce ecosystems, is a critical task for accurately identifying and matching similar product offers across the web. Traditional approaches to entity resolution have primarily relied on textual information, but the increasing availability of diverse data modalities has led to the adoption of a multimodal approach. This paper introduces an innovative intermediate fusion architecture for multimodal product matching, effectively combining textual information from RoBERTa embeddings and visual information from Swin-Transformer embeddings. Our approach enhances matching accuracy by leveraging the complementary nature of text and image modalities. Experimental results on the WDC Shoes and Zalando datasets show the superiority of our proposed approach compared to unimodal models and multimodal baselines. The outcomes highlight the potential for multimodal product matching to improve entity resolution in online marketplaces, thereby enhancing the user shopping experience.

Keywords

entity resolution, product matching, multimodal data, intermediate fusion

1. Introduction

Web-based entity resolution, particularly in product matching, serves as a cornerstone for online marketplaces, aiding in price comparison, reducing duplicate listings, and ensuring a seamless shopping experience. However, traditional approaches relying solely on textual data encounter limitations due to ambiguous or incomplete descriptions and the vast expanse of online marketplaces. In response, the adoption of a multimodal approach, integrating both textual and visual data, has gained traction to enhance matching accuracy.

Multimodal product matching presents a unique opportunity to leverage the complementary nature of textual descriptions and visual images. While textual data provides context and details about products, visual information captures nuances and attributes that are challenging to convey through text alone. This fusion of modalities holds promise for achieving more precise and reliable matches.

Fusion strategies play a pivotal role in multimodal product matching, determining how information from different modalities is integrated. Early fusion combines raw data from text and images at the onset, potentially losing modality-specific details. Late fusion defers integration until each modality has been independently processed, possibly missing subtle inter-modal relation-

ships. Intermediate fusion strikes a balance by merging modality-specific features at an intermediate stage, allowing for a nuanced integration of complementary information while retaining modality-specific details.

This paper focuses on the concept of intermediate fusion, where textual and visual features are combined at an intermediate representation level. We make the following contributions:

- We introduce a publicly available, high quality, and challenging benchmark dataset for multimodal product matching.
- We propose an intermediate fusion architecture combining textual information from RoBERTa [1] and visual information from Swin-Transformer embeddings [2].
- We comprehensively evaluate our intermediate fusion approach against a range of unimodal and multimodal approaches.

The remainder of this paper is structured as follows. In Section 2 we discuss related work. We introduce the benchmark datasets that we utilize to evaluate the effectiveness of our proposed intermediate fusion architecture in Section 3. Section 4 introduces our intermediate fusion architecture. Subsequently, we present our evaluation results in Section 5. Lastly, Section 6 summarizes our findings and outlines potential avenues for further research in this field. The full code implementation, data, and accompanying documentation can be accessed at the following repository: <https://git.informatik.uni-leipzig.de/jp31zusu/intermediate-fusion-for-multimodal-product-matching>.

35th GI-Workshop on Foundations of Databases, May 22-24, 2024, Herdecke, Germany

*Corresponding author.


✉ pollack@informatik.uni-leipzig.de (J. Pollack);

koepcke@hs-mittweida.de (H. Köpcke);

rahm@informatik.uni-leipzig.de (E. Rahm)

ORCID: [0000-0003-2501-2609](https://orcid.org/0000-0003-2501-2609) (H. Köpcke); [0000-0002-2665-1114](https://orcid.org/0000-0002-2665-1114)

(E. Rahm)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Related work

The field of Entity Resolution (ER) has been a subject of active research since the 1950s [3]. To gain an in-depth understanding of ER, readers are directed to recent books and surveys [4, 5, 6, 7]. An overview of the (increasing) usage of neural networks and deep learning for entity resolution can be found in [8]. The combination of text and image data in deep learning systems is called multimodal deep learning. An overview on tasks, datasets and problems in this new field can be found in [9].

Applying ER on e-commerce data has been explored with various approaches and methods within the last years. The initial approaches for product matching utilize only textual information and supervised machine learning to categorize and match products [10, 11, 12, 13]. An early approach to also consider image data for product matching is [14] but with the goal of enriching textual product descriptions with images (the images are only in one of the two datasets).

Recently several multimodal product matching approaches have emerged, employing different fusion strategies. It's worth noting that the terminology and definitions of these fusion strategies can vary considerably across the literature [15, 16]. To ensure clarity and consistency, we establish the following definitions:

- In Early Fusion, data from different modalities are directly combined. An example would be overlaying text onto an image and then processing both the text and image simultaneously using a CNN.
- In Intermediate Fusion, features from different modalities are processed separately, combined, and then further processed before a final decision is made. This means that after combining the features, additional processing steps, such as passing them through a neural network or applying further transformations, are carried out before the final decision-making.
- In Late Fusion, features from different modalities are still processed separately, but a decision is made directly after the fusion step, without further processing of the combined features. This means that once the features are combined, a decision-making mechanism, such as applying a threshold or weighted averaging, is immediately applied to determine the final output.
- Hybrid Fusion integrates modalities at multiple stages within the processing pipeline. For example, combining text and image features at an intermediate stage, then further processing them, and finally combining the results again with those from unimodal methods.

The key difference between Intermediate Fusion and Late Fusion lies in the timing of decision-making relative to

the fusion process. Intermediate Fusion involves additional processing steps after feature fusion, while Late Fusion makes the decision directly after fusion without additional processing.

With these clarified definitions in place, we proceed to categorize existing multimodal product matching approaches according to these distinct fusion strategies.

Wilke et al. [17] apply late fusion enhancing the Deep-Matcher framework [18]. They use FastText for token-level text embeddings, which are then aggregated via Recurrent Neural Network (RNN). For images, they employ a pre-trained Residual Neural Network (ResNet) with a fully connected layer. Similarity is computed with Euclidean distances and concatenation, followed by classification using two fully connected layers.

Ali Mazhar et al. [19] propose two intermediate fusion architectures for multimodal networks, one using element-wise multiplication to combine pre-trained image embeddings (ResNet, VGG, or MobileNet) with text embeddings from a character-level CNN model, followed by logistic regression for classification, and the other relying on bidirectional triplet loss.

Gupte et al. [20] propose a weighted hybrid fusion strategy, combining BERT text embeddings and ResNet image embeddings in Siamese Networks. Their approach outperforms text-based and late fusion methods.

Das et al. [21] employ intermediate fusion, prioritizing product images as their primary data source, using title information to emphasize relevant image regions. Their model consists of a global branch for overall image feature extraction, a local branch for specific feature derivation, and incorporates self-attention mechanisms. While showing promise on a multi-category dataset, it exhibits potential performance variations in low-data scenarios.

Valenciano et al. [22] delve into both intermediate fusion and late fusion strategies. The first method integrates modalities early, combining ResNet image embeddings with joint BERT text embeddings, followed by BiLSTM and hybrid pooling. The second method utilizes a Siamese Network, computing Euclidean distances between ResNet image embeddings, and later concatenating the outputs with joint BERT text embedding. The first strategy outperforms the second, indicating that relying solely on Euclidean distances of images may not be effective for multimodal matching.

Feng et al. [23] take a late fusion approach, integrating text embeddings generated by Robustly Optimized BERT Pre-training Approach (RoBERTa) [1] models with image embeddings from Hierarchical Vision Transformer (Swin-Transformer) [2] models, leveraging a K-gram Exponential decay scheme for text embeddings. Notably, they opt for Swin-Transformers instead of ResNets for image embeddings, capitalizing on self-attention mechanisms for comprehensive context capture.

Early fusion can lead to larger input vectors with redundancies, causing increased computational complexity [24], while late fusion may struggle to capture complex inter-modal relationships, limiting effectiveness in multi-modal product matching [20]. Therefore, we employ an intermediate fusion approach, inspired by the methodologies of Valenciano et al. [22] and Feng et al. [23]. Our method enhances text embeddings with a BiLSTM layer and hybrid pooling, while also integrating higher-level image embeddings from Siamese Networks in an intermediate training step. Notably, we depart from the conventional use of ResNets for image embeddings, opting for Swin-Transformers, as demonstrated in the work of Feng et al. [23].

3. Benchmark datasets

In the realm of research on multimodal product matching, one of the primary challenges has been the limited availability of publicly accessible datasets containing ground-truth information for product matching tasks [17].

To address this challenge, Primpeli et al. [25] generously shared their WDC Training Dataset and Gold Standard for Large-scale Product Matching with the scientific community. Originally designed for text-based unimodal product matching, these datasets have been instrumental in advancing the field. Building upon this foundation, Wilke et al. [17] undertook the task of augmenting the WDC product matching dataset with images sourced from the internet. They meticulously verified a subset of products in the shoe category from the original datasets to create multimodal datasets. In these datasets, each product is associated with an image, and no erroneous or distracting images are included.

More recently, Lamm et al. [26] introduced a publicly available large-scale dataset for visual entity matching, based on a real-world use case within the retail domain. Unfortunately, this dataset does not provide textual information for the products, rendering it unsuitable for the evaluation of multimodal approaches that rely on both text and images.

While the WDC Shoes dataset is primarily centered on footwear, we introduce an additional dataset that encompasses a wide range of clothing items. This dataset is constructed using product titles and images extracted from three online stores: Zalando, Tommy Hilfiger, and Gerry Weber. We crawled all available products from the Tommy Hilfiger and Gerry Weber online stores. When collecting data from Zalando, our focus was on products from the Tommy Hilfiger and Gerry Weber brands. Establishing ground truth for this diverse dataset relied on a semi-supervised approach, combining manufacturer part (MPN) matching with manual inspection.

In order to create a challenging dataset that can im-

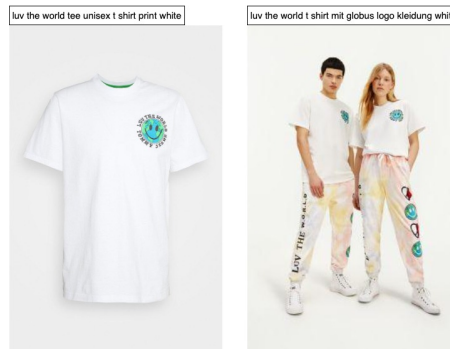


Figure 1: Hard positive pair from Zalando dataset



Figure 2: Hard negative pair from Zalando dataset

prove model performance, we curated non-matching (negative) pairs. To accomplish this, we leveraged text and image embeddings generated by pre-trained RoBERTa and Swin-Transformer models, incorporating them into multimodal embeddings. Subsequently, we employed an approximate nearest neighbor search within the resulting embedding space to identify, for each Tommy Hilfiger and Gerry Weber product, the most similar non-matching Zalando product. As a result, our negative samples included products that exhibited visual similarities in their images or shared textual characteristics but ultimately did not constitute a match. Exemplary matching and non-matching pairs represented in terms of their pre-processed product titles and image are shown in Figures 1 and 2.

We divided the resulting pairs into both training and test datasets, with a specific condition: the Zalando products in both sets of pairs were mutually exclusive. This separation allowed for a more accurate assessment of model performance on previously unseen data. For a quantitative summary of the datasets, refer to Table 1

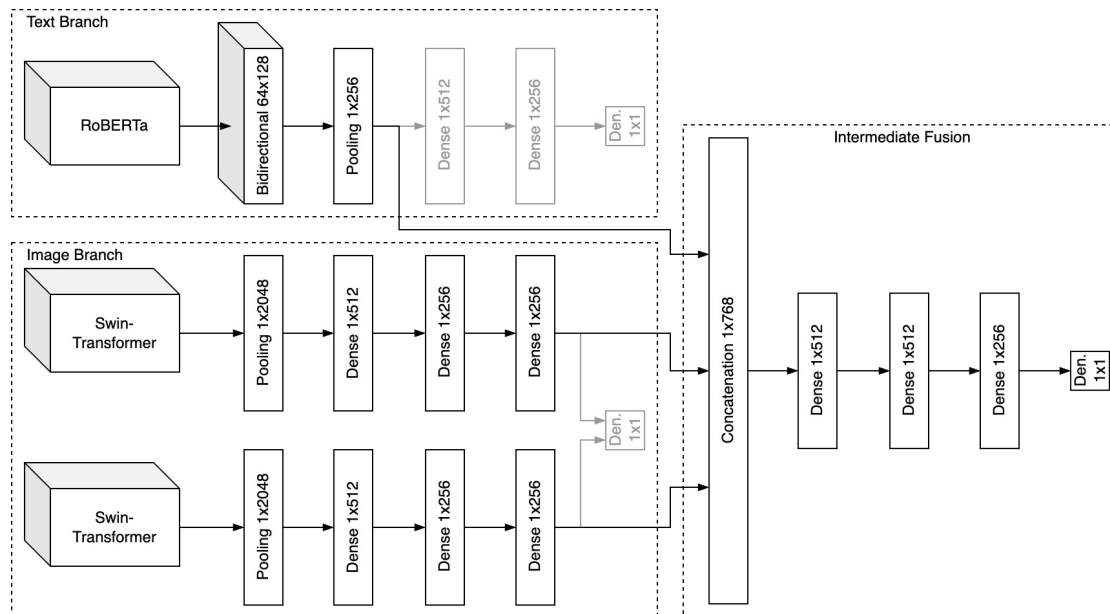


Figure 3: Intermediate Fusion architecture

Table 1
Quantitative overview of datasets

Dataset	Products	Matches	Non-matches
WDC Shoes train	950	1350	6286
WDC Shoes test	813	206	586
Zalando train	6676	945	3843
Zalando test	1431	178	785

which outlines the total unique products, matching (positive) pairs, and non-matching (negative) pairs for each dataset.

4. Intermediate Fusion Architecture

Our intermediate fusion architecture strategically integrates text-based information and image-based information, harnessing the unique strengths of both modalities and effectively capturing intricate relationships between them.

We utilize RoBERTa embeddings to ensure a robust representation of textual information, backed by their exceptional performance across various natural language processing tasks, including multimodal product matching [23].

We utilize the Swin-Transformer architecture for visual information representation, known for its efficiency in capturing spatial dependencies and long-range de-

pendencies in various image processing tasks, including multimodal product matching [23]. Swin-Transformer, a recent advancement in vision transformer architecture, employs shifted windows and hierarchical transformer blocks to replace traditional convolutional layers, facilitating parallel processing and superior extraction of visual features, contextual information, and semantic relationships within images compared to established techniques like ResNet.

Our multimodal model combines RoBERTa-based text information with Swin-Transformer-based image data using an intermediate fusion strategy. We chose intermediate fusion over early or late fusion because it enables nuanced integration of high-level representations from different modalities, preserving their distinctive characteristics and adapting to the complexities of multimodal data, as empirically supported by prior research [20, 22]. We build upon Valenciano et al.’s multimodal models, making specific modifications to address their limitations. While Valenciano et al. found Siamese Network-based models superior in unimodal image contexts, the concatenation of Euclidean distances with text embeddings did not improve performance in multimodal settings. Consequently, for better multimodal performance, they resorted to direct classification. In this work, we introduce an intermediate training step designed to seamlessly incorporate higher-level image embeddings obtained from Siamese Networks. These embeddings contain rich information as Siamese Networks inherently learn to identify

similarities and differences between inputs, especially those critical for making matching decisions. Our architecture is depicted in Figure 3 and comprises two primary branches: a text branch and an image branch. Each branch undergoes individual training in an intermediate step before subsequent fully connected layers carry out the final classification.

In the text branch, input texts are structured as sequences in the format "[CLS] text 1 [SEP] text 2 [SEP]" to leverage RoBERTa’s ability to handle text pairs. The output of the [CLS] token encapsulates the sentence-level embedding for the entire text pair. This embedding is refined through a Bidirectional LSTM (BiLSTM) layer that conducts bidirectional analysis, encompassing both preceding and subsequent context. The BiLSTM layer’s outputs are then channeled through a hybrid pooling layer, amalgamating max-pooling and mean-pooling to capture key and average representations, respectively. This process transforms the embeddings into one-dimensional vectors. During the intermediate training step, these outputs from the hybrid pooling layer are further processed by multiple fully connected layers for classification.

In the image branch, Swin-Transformer embeddings are fed into an average pooling layer, transforming them into one-dimensional vectors, which are then fine-tuned using multiple fully connected layers. Adopting a Siamese Network approach, both sub-networks share the same parameters. The Euclidean distance between the resulting embeddings of both sub-networks is employed during the intermediate training step for classification.

The outputs of both sub-networks in the image branch serve as higher-level image embeddings, which are concatenated with the fine-tuned BERT embedding from the text branch. This combined representation is then passed through multiple fully connected layers. Our architecture is meticulously designed to capitalize on information from both textual and visual modalities, effectively capturing the intricate relationships between them.

5. Evaluation

We evaluate the performance of our proposed intermediate fusion approach on two distinct datasets: WDC Shoes and Zalando product matching datasets. Our evaluation entails a comprehensive comparison against a range of unimodal and multimodal baseline models, with Table 2 offering a detailed summary of the experimental outcomes. In order to provide a thorough evaluation, we report the F1-score, precision, and recall metrics for each case. The reported results are an average computed over three training and test runs to ensure robust estimates of performance, effectively mitigating the impact of stochasticity.

The outcomes show that our text-based model consis-

tently outperforms its image-based counterpart across both datasets. This observation underscores the potent role of textual data in product matching, aligning with the textual richness often present in product descriptions and titles. However, our central contribution lies in the amalgamation of both unimodal models into a multimodal model (Intermediate Fusion), resulting in a significant enhancement in performance.

It’s worth noting that the ability of our models to effectively harness complementary information from each modality proves especially advantageous when confronted with ambiguous or incomplete data. By intelligently fusing textual and visual features, the model demonstrates a capacity to compensate for missing information, promoting more accurate and robust product matching. Our multimodal model not only achieved the highest F1-score but also demonstrates a more balanced trade-off between precision and recall, which is often desirable in real-world applications.

We compared our intermediate fusion approach with multimodal models by Wilke et al. [17] and Valenciano et al. [22]. Our model showed significant performance improvements over Wilke et al.’s approach, mainly due to our effective fusion strategy. Additionally, we re-implemented Valenciano et al.’s ImageBERT model and its unimodal counterparts [22], further highlighting the performance gains achieved by our intermediate fusion architecture, primarily attributed to our strategic choice of Swin-Transformer embeddings and enhanced fusion strategy.

The Zalando dataset demonstrated lower performance than the WDC Shoes dataset, attributed to greater variability in both textual descriptions and visual attributes among matching products. Additionally, the WDC Shoes dataset features overlapping products between training and testing sets, potentially biasing model evaluations by assessing performance on partially seen data. Moreover, the WDC Shoes test set has a higher proportion of matching pairs compared to non-matching ones, which may not accurately mirror real-world scenarios where non-matching products typically outnumber matching ones. Thus, the Zalando dataset provides a more realistic assessment of model performance in practical settings.

Products in the Zalando dataset varied more in both textual descriptions and visual representations, making it challenging for models to establish consistent relationships, especially with limited training data. Images of matching pairs in the Zalando dataset could differ significantly, presenting varied contexts like different clothing combinations (See Figure 1). Similarly, non-matching product images could appear very similar, making accurate determination challenging without additional context (See Figure 2). Some non-matching pairs differed only in minor variants, leading to nearly identical images and text descriptions, further complicating differentia-

Table 2

Comparison of experimental results. The experimental results for DeepMatcher are obtained from Wilke et al. [17], while the experimental results for Valenciano et al.’s ImageBERT [22] were computed using a custom reimplementaion.

Dataset	Model	F1-Score	Precision	Recall
WDC Shoes	Unimodal			
	Our Text Branch (RoBERTa)	0.8477	0.8102	0.8900
	DeepMatcher title (FastText) [17]	0.8520	0.7930	0.9190
	Reimpl. BERT-based (BERT) [22]	0.7520	0.6409	0.9288
	Our Image Branch (Swin-Transformer)	0.7593	0.7058	0.8223
	DeepMatcher image (ResNet) [17]	0.7310	0.6120	0.9080
	Reimpl. 2-CNN (ResNet) [22]	0.5743	0.4713	0.7395
	Multimodal			
	Our Intermediate Fusion	0.8829	0.8704	0.8968
	DeepMatcher title+image (Late Fusion) [17]	0.8560	0.7920	0.9300
Reimpl. ImageBERT (Intermediate Fusion) [22]	0.7428	0.6222	0.9223	
Zalando	Unimodal			
	Our Text Branch (RoBERTa)	0.6230	0.4982	0.8464
	Reimpl. BERT-based (BERT) [22]	0.5662	0.4137	0.8970
	Our Image Branch (Swin-Transformer)	0.6178	0.6206	0.6161
	Reimpl. 2-CNN (ResNet) [22]	0.4807	0.3660	0.7153
	Multimodal			
	Our Intermediate Fusion	0.7401	0.7144	0.7697
Reimpl. ImageBERT (Intermediate Fusion) [22]	0.6048	0.4624	0.8858	

tion. Despite these challenges, our multimodal intermediate fusion approach notably improved performance compared to unimodal methods.

6. Conclusion and Future work

In this paper, we propose an intermediate fusion architecture for multimodal product matching, enhancing web-based entity resolution. Our model effectively captures complex relationships between product descriptions and images by merging fine-tuned RoBERTa textual embeddings with higher-level Swin-Transformer visual embeddings from Siamese Networks. Experimental results on WDC Shoes and Zalando datasets demonstrate the superiority of our approach over unimodal models, highlighting its potential for enhancing entity resolution in online marketplaces and e-commerce. Future work will focus on scaling the approach, incorporating advanced data augmentation techniques, and exploring cross-modal attention mechanisms for improved performance.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by the Sächsische Staatsministerium für Wissenschaft Kultur und Tourismus in the program Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification: ScaDS.AI. Computations for this work were done (in part) using resources of the Leipzig University Computing Centre.

References

- [1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019).
- [2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 9992–10002.
- [3] H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James, Automatic linkage of vital records, Science 130 (1959) 954–959.
- [4] O. Binette, R. C. Steorts, (almost) all of entity resolution, Science Advances 8 (2022) eabi8021.
- [5] P. Christen, T. Ranbaduge, R. Schnell, Linking Sensitive Data - Methods and Techniques for Practical Privacy-Preserving Information Sharing, Springer, 2020.
- [6] X. L. Dong, D. Srivastava, Big Data Integration, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2015.
- [7] G. Papadakis, E. Ioannou, E. Thanos, T. Palpanas, The Four Generations of Entity Resolution, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2021.
- [8] N. Barlaug, J. A. Gulla, Neural networks for entity matching: A survey, ACM Trans. Knowl. Discov. Data 15 (2021) 52:1–52:37.
- [9] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods, J. Artif.

- Intell. Res. 71 (2021) 1183–1317.
- [10] H. Köpcke, A. Thor, S. Thomas, E. Rahm, Tailoring entity resolution for matching product offers, in: Proc. 15th Int. Conf. on Extending Database Technology (EDBT), 2012, pp. 545–550.
- [11] C. d’Amato, P. Ristoski, P. Petrovski, P. Mika, H. Paulheim, A machine learning approach for product matching and categorization, *Semant. Web* 9 (2018) 707–728.
- [12] J. Li, Z. Dou, Y. Zhu, X. Zuo, J. Wen, Deep cross-platform product matching in e-commerce, *Inf. Retr. J.* 23 (2020) 136–158.
- [13] R. Peeters, C. Bizer, Supervised contrastive learning for product matching, in: Companion Proceedings of the Web Conference 2022, WWW ’22, Association for Computing Machinery, New York, NY, USA, 2022, p. 248–251.
- [14] P. Ristoski, P. Petrovski, P. Mika, H. Paulheim, A machine learning approach for product matching and categorization, *Semantic Web* 9 (2018) 707–728.
- [15] S. R. Stahlschmidt, B. Ulfenborg, J. Synnergren, Multimodal deep learning for biomedical data fusion: a review, *Briefings in Bioinformatics* 23 (2022) bbab569.
- [16] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, *Information fusion* 37 (2017) 98–125.
- [17] M. Wilke, E. Rahm, Towards multi-modal entity resolution for product matching, in: GvDB, 2021.
- [18] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, Deep learning for entity matching: A design space exploration, *Proceedings of the 2018 International Conference on Management of Data* (2018).
- [19] K. A. Mazhar, M. Brodtbeck, G. Gühring, Similarity learning of product descriptions and images using multimodal neural networks, *Natural Language Processing Journal* 4 (2023) 100029.
- [20] K. Gupte, L. X. Pang, H. Vuyyuri, S. Pasumarty, Multimodal product matching and category mapping: Text+image based deep neural network, *2021 IEEE International Conference on Big Data (Big Data)* (2021) 4500–4505.
- [21] N. Das, A. Joshi, P. Yenigalla, G. Agrwal, Maps: Multimodal attention for product similarity, *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022) 2988–2996.
- [22] R. Estrada-Valenciano, V. Muñoz-Sánchez, H. D. la Torre-Gutiérrez, An entity-matching system based on multimodal data for two major e-commerce stores in Mexico, *Mathematics* (2022).
- [23] C.-F. Feng, W. Chen, C. Chen, T. Xu, E. Chen, Multimodal representation learning-based product matching, in: *China Conference on Knowledge Graph and Semantic Computing*, 2022.
- [24] D. Ramachandram, G. W. Taylor, Deep multimodal learning: A survey on recent advances and trends, *IEEE Signal Processing Magazine* 34 (2017) 96–108.
- [25] A. Primpeli, R. Peeters, C. Bizer, The wdc training dataset and gold standard for large-scale product matching, *Companion Proceedings of The 2019 World Wide Web Conference* (2019).
- [26] B. Lamm, J. Keuper, Retail-786k: a large-scale dataset for visual entity matching, 2023.