Extending CLIP for Category-to-image Retrieval in E-commerce

Mariya Hendriksen [1], Maurits Bleeker [1], Svitlana Vakulenko [1,2], Nanne van Noord [1], Ernst Kuiper [3], Maarten de Rijke [1]

April 12, 2022

[1]University of Amsterdam [2]Amazon (now) [3]bol.com

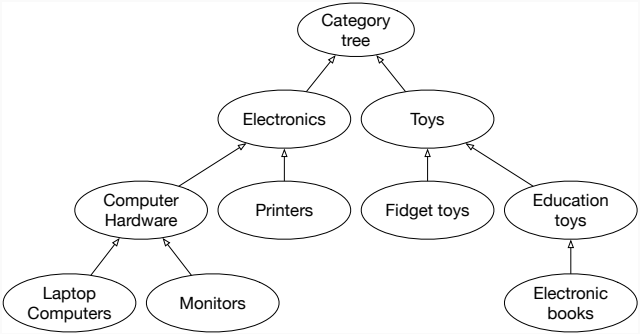
Introduction

Motivation: product category tree

Assist customers when navigating product catalogue.

Ability to retrieve an image for a given category is a challenge due to:

- noisy category and product data
- size and dynamic character of product catalogues



Motivation: multi-modal data in e-commerce

- Current e-commerce search focuses on textual and behavioural signals.
- Multimodal product data is barely used.
- Prior work mainly on Fashion retrieval.
- Knowledge gap: multimodal retrieval in general e-commerce domain.

Category-to-image (CtI) retrieval task

Task

Given a *category* and a *collection of products*, retrieve a list of images of products that belong to a given category.

Task key characteristics:

- We operate on an e-commerce category tree $\longrightarrow$ categories vary in granularity, e.g., "Home & Living" $\rightarrow$ "Kitchen" $\rightarrow$ "Coffee Machine".
- The category tree is not fixed $\longrightarrow$ aim to generalize towards unseen categories.
- Multimodal product information, i.e., textual, visual, and attribute information.

Category-to-image (CtI) retrieval task

Task

Given a *category* and a *collection of products*, retrieve a list of images of products that belong to a given category.

Task key characteristics:

- We operate on an e-commerce category tree $\longrightarrow$ categories vary in granularity, e.g., "Home & Living" $\rightarrow$ "Kitchen" $\rightarrow$ "Coffee Machine".
- The category tree is not fixed $\longrightarrow$ aim to generalize towards unseen categories.
- Multimodal product information, i.e., textual, visual, and attribute information.

Research questions

1. How do unimodal vs. multimodal models perform on the task and how does the performance differ w.r.t. category granularity?
2. How does different combinations of multimodal product information impacts the performance on the task?

Research questions

1. How do unimodal vs. multimodal models perform on the task and how does the performance differ w.r.t. category granularity?
2. How does different combinations of multimodal product information impacts the performance on the task?

Approach

Task overview

Definition

- *Input*: category-product pairs.
- *Query*: category.
- *Return*: a ranked list of images that belong to the category.

Metrics

Precision@K where $K = \{1, 5, 10\}$, mAP@K where $K = \{5, 10\}$, and R-precision.

Task overview

Definition

- *Input*: category-product pairs.
- *Query*: category.
- *Return*: a ranked list of images that belong to the category.

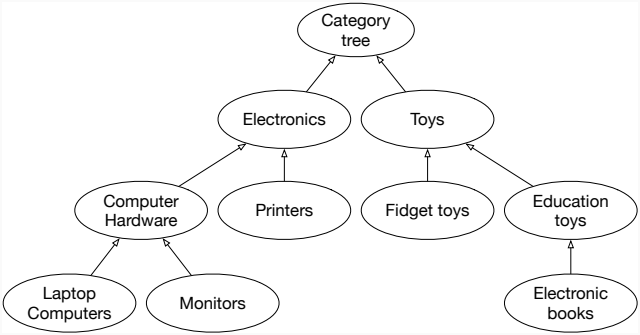
Metrics

Precision@K where $K = \{1, 5, 10\}$, mAP@K where $K = \{5, 10\}$, and R-precision.

Evaluation w.r.t. category granularity

For every category-product pair, we sample a category in one of three settings:

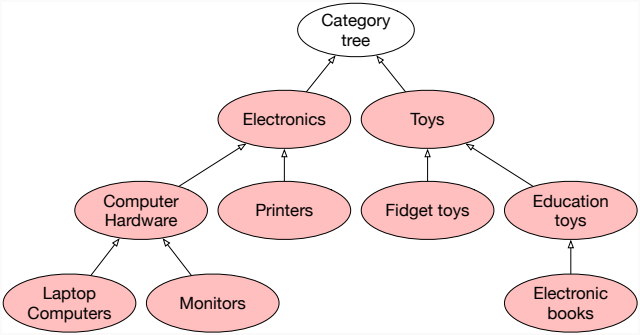
- All categories
- Most general category
- Most specific category



Evaluation w.r.t. category granularity

For every category-product pair, we sample a category in one of three settings:

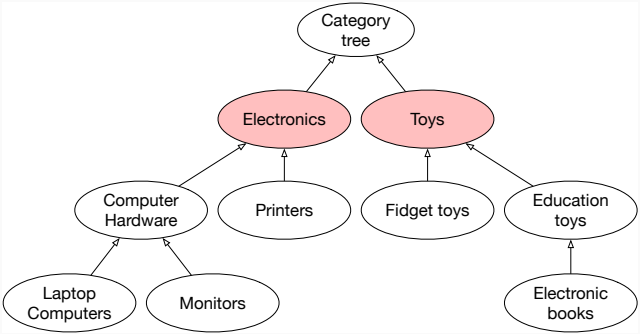
- All categories
- Most general category
- Most specific category



Evaluation w.r.t. category granularity

For every category-product pair, we sample a category in one of three settings:

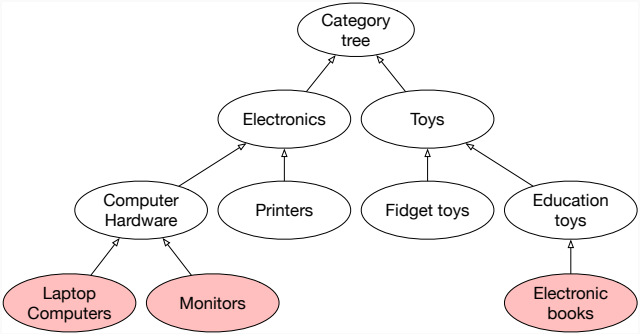
- All categories
- Most general category
- Most specific category



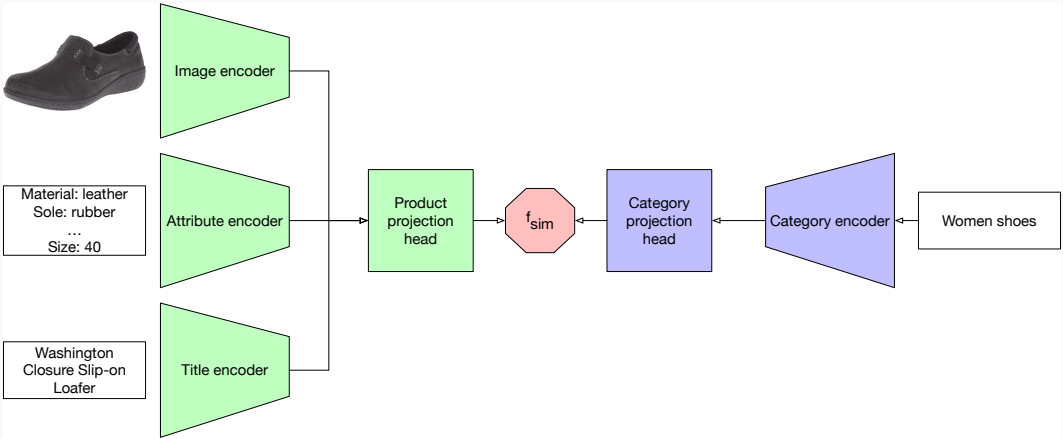
Evaluation w.r.t. category granularity

For every category-product pair, we sample a category in one of three settings:

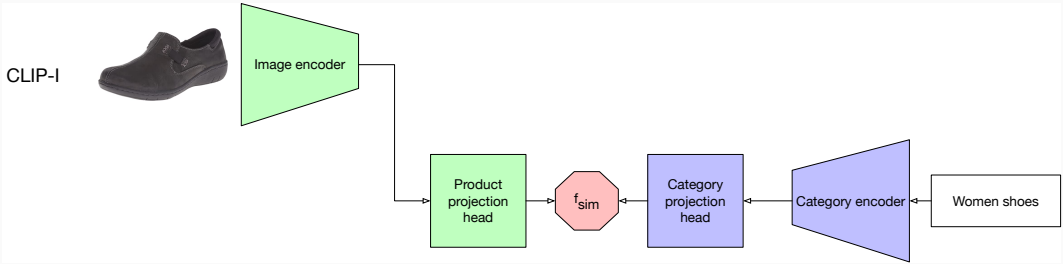
- All categories
- Most general category
- Most specific category



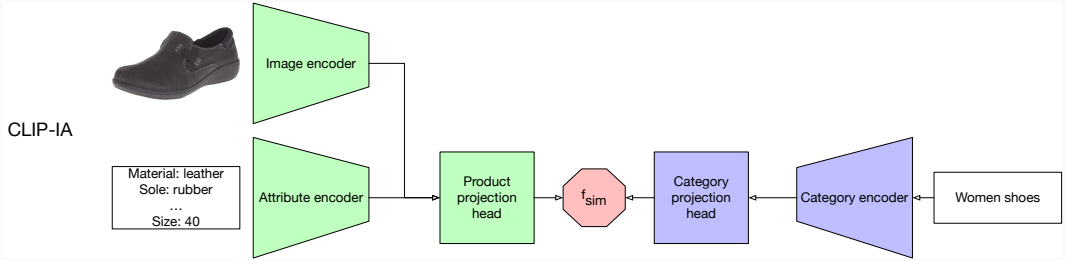
Model Overview



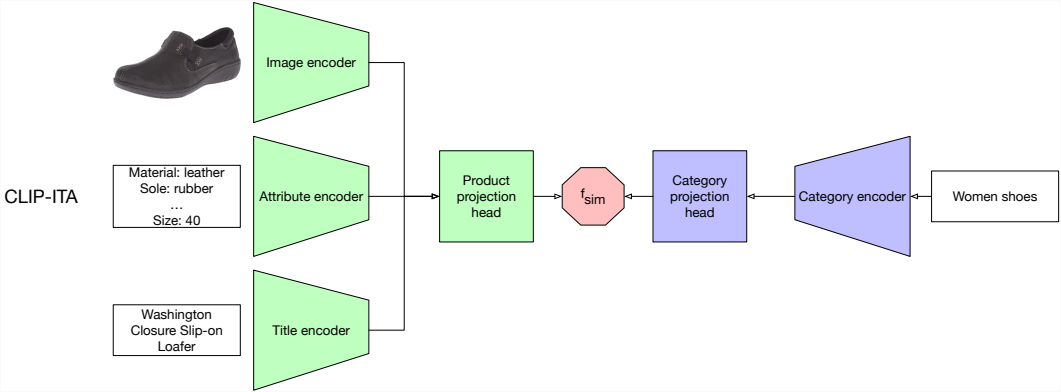
Model Overview



Model Overview



Model Overview



Model Overview

Image encoder

Material: leather
Sole: rubber
...
Size: 40

Attribute encoder

Washington
Closure Slip-on
Loafer

Title encoder

Product projection head

f_{sim}

Category projection head

Category encoder

Women shoes

$$\mathcal{L} = \frac{1}{\beta} \sum_{j=1}^{\beta} \left(\lambda \ell_j^{(p \to c)} + (1 - \lambda) \ell_j^{(c \to p)} \right)$$

Bidirectional contrastive loss

Comprises two losses:

1. Category-to-product loss:

$$\ell_j^{(c \to p)} = -\log \frac{\exp(f_{sim}(\mathbf{c}_j, \mathbf{p}_j)/\tau)}{\sum_{k=1}^{\beta} \exp(f_{sim}(\mathbf{c}_j, \mathbf{p}_k)/\tau)}, \tag{1}$$

2. Product-to-category loss:

$$\ell_j^{(p \to c)} = -\log \frac{\exp(f_{sim}(\mathbf{p}_j, \mathbf{c}_j)/\tau)}{\sum_{k=1}^{\beta} \exp(f_{sim}(\mathbf{p}_j, \mathbf{c}_k)/\tau)}. \tag{2}$$

Experiments

Experimental Setup

Dataset

- Amazon XMarket dataset [1]
- Textual, visual, attribute information, category tree

Baselines

- Text-only: BM25, and MPNet
- Multimodal: CLIP

Experimental Setup

Dataset

- Amazon XMarket dataset [1]
- Textual, visual, attribute information, category tree

Baselines

- Text-only: BM25, and MPNet
- Multimodal: CLIP

Evaluation

1. Baselines, BM25, CLIP, MPNet.
2. Image-based product representations, CLIP-I.
3. Image and attribute-based product representations, CLIP-IA.
4. Image, attribute, and title-based product representations, CLIP-ITA.

Evaluation

1. Baselines, BM25, CLIP, MPNet.
2. Image-based product representations, CLIP-I.
3. Image and attribute-based product representations, CLIP-IA.
4. Image, attribute, and title-based product representations, CLIP-ITA.

Evaluation

1. Baselines, BM25, CLIP, MPNet.
2. Image-based product representations, CLIP-I.
3. Image and attribute-based product representations, CLIP-IA.
4. Image, attribute, and title-based product representations, CLIP-ITA.

Evaluation

1. Baselines, BM25, CLIP, MPNet.
2. Image-based product representations, CLIP-I.
3. Image and attribute-based product representations, CLIP-IA.
4. Image, attribute, and title-based product representations, CLIP-ITA.

Evaluation of baselines

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			All categories			
BM25	0.01	0.01	0.01	0.01	0.01	0.01
CLIP	0.01	0.02	0.02	0.03	0.04	0.02
MPNet	0.01	**0.06**	**0.06**	**0.07**	**0.09**	**0.05**

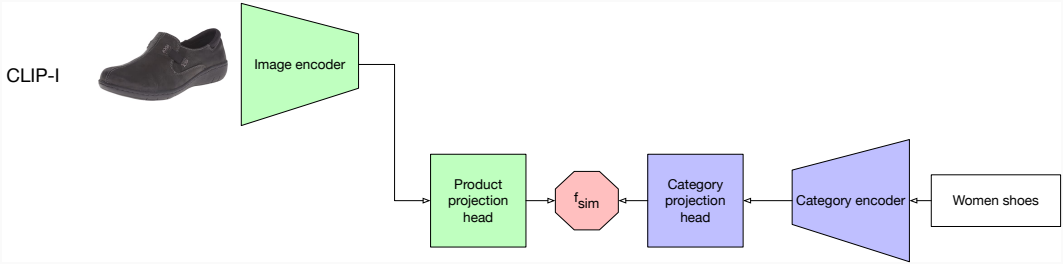
Evaluation of baselines

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
	Most general category					
BM25	2.94	4.71	4.71	8.33	8.28	4.48
CLIP	11.76	12.35	11.76	16.12	15.18	**9.47**
MPNet	**14.70**	**15.8**	**15.01**	**18.44**	**18.78**	9.35

Evaluation of baselines

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			Most specific category			
BM25	0.02	0.02	0.01	0.01	0.01	0.01
CLIP	11.92	9.81	9.23	15.12	14.95	8.14
MPNet	**33.36**	**28.56**	**26.93**	**37.43**	**36.77**	**25.29**

Evaluation of image-based product representations, CLIP-I



Evaluation of image-based product representations, CLIP-I

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
	All categories					
BM25	0.01	0.01	0.01	0.01	0.01	0.01
CLIP	0.01	0.02	0.02	0.03	0.04	0.02
MPNet	0.01	0.06	0.06	0.07	0.09	0.05
CLIP-I	**3.3**	**3.8**	**3.79**	**6.81**	**7.25**	**3.67**

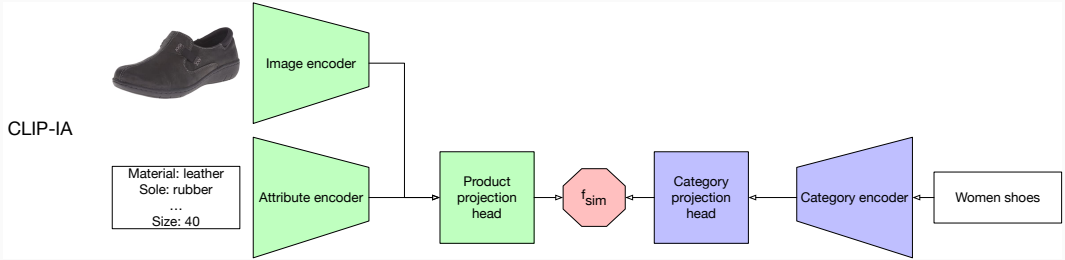
Evaluation of image-based product representations, CLIP-I

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			Most general category			
BM25	2.94	4.71	4.71	8.33	8.28	4.48
CLIP	11.76	12.35	11.76	16.12	15.18	9.47
MPNet	14.70	15.8	15.01	18.44	18.78	9.35
CLIP-I	**17.85**	**17.14**	**16.78**	**19.88**	**20.14**	**13.02**

Evaluation of image-based product representations, CLIP-I

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
	Most specific category					
BM25	0.02	0.02	0.01	0.01	0.01	0.01
CLIP	11.92	9.81	9.23	15.12	14.95	8.14
MPNet	**33.36**	**28.56**	**26.93**	**37.43**	**36.77**	**25.29**
CLIP-I	14.06	12.11	11.53	18.24	17.9	11.22

Evaluation of image and attribute-based product representations, CLIP-IA



Evaluation of image and attribute-based product representations, CLIP-IA

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			All categories			
BM25	0.01	0.01	0.01	0.01	0.01	0.01
CLIP	0.01	0.02	0.02	0.03	0.04	0.02
MPNet	0.01	0.06	0.06	0.07	0.09	0.05
CLIP-I	**3.3**	**3.8**	**3.79**	**6.81**	**7.25**	**3.67**
CLIP-IA	2.5	3.34	3.29	5.95	6.24	3.27

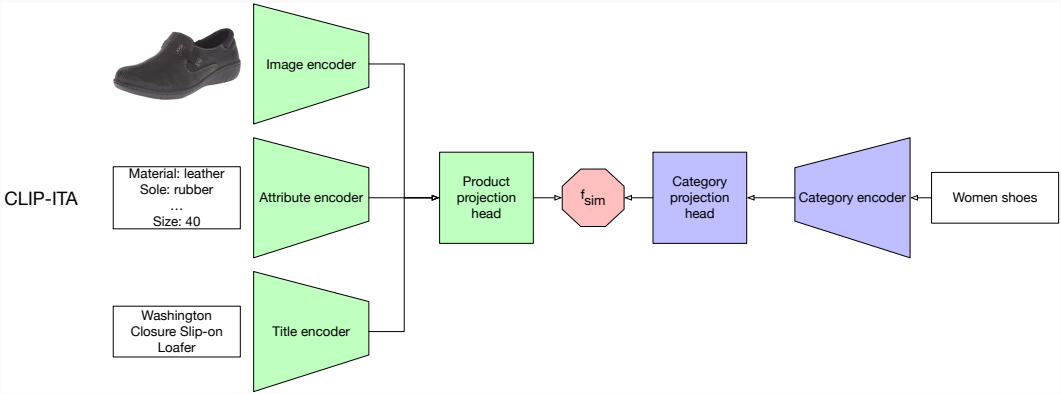
Evaluation of image and attribute-based product representations, CLIP-IA

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			Most general category			
BM25	2.94	4.71	4.71	8.33	8.28	4.48
CLIP	11.76	12.35	11.76	16.12	15.18	9.47
MPNet	14.70	15.8	15.01	18.44	18.78	9.35
CLIP-I	17.85	17.14	16.78	19.88	20.14	13.02
CLIP-IA	**21.42**	**21.91**	**22.78**	**25.59**	**26.29**	**20.74**

Evaluation of image and attribute-based product representations, CLIP-IA

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			Most specific category			
BM25	0.02	0.02	0.01	0.01	0.01	0.01
CLIP	11.92	9.81	9.23	15.12	14.95	8.14
MPNet	33.36	28.56	26.93	37.43	36.77	25.29
CLIP-I	14.06	12.11	11.53	18.24	17.9	11.22
CLIP-IA	**35.3**	**30.21**	**29.32**	**39.93**	**39.27**	**28.86**

Evaluation of image, attribute, and title-based product representations, CLIP-ITA



Evaluation of image, attribute, and title-based product representations, CLIP-ITA

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			All categories			
BM25	0.01	0.01	0.01	0.01	0.01	0.01
CLIP	0.01	0.02	0.02	0.03	0.04	0.02
MPNet	0.01	0.06	0.06	0.07	0.09	0.05
CLIP-I	3.3	3.8	3.79	6.81	7.25	3.67
CLIP-IA	2.5	3.34	3.29	5.95	6.24	3.27
CLIP-ITA	**9.9**	**13.27**	**13.43**	**20.3**	**20.53**	**13.42**

Evaluation of image, attribute, and title-based product representations, CLIP-ITA

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
			Most general category			
BM25	2.94	4.71	4.71	8.33	8.28	4.48
CLIP	11.76	12.35	11.76	16.12	15.18	9.47
MPNet	14.70	15.8	15.01	18.44	18.78	9.35
CLIP-I	17.85	17.14	16.78	19.88	20.14	13.02
CLIP-IA	21.42	21.91	22.78	25.59	26.29	20.74
CLIP-ITA	**35.71**	**30.95**	**30.95**	**35.51**	**34.28**	**25.79**

Evaluation of image, attribute, and title-based product representations, CLIP-ITA

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
	Most specific category					
BM25	0.02	0.02	0.01	0.01	0.01	0.01
CLIP	11.92	9.81	9.23	15.12	14.95	8.14
MPNet	33.36	28.56	26.93	37.43	36.77	25.29
CLIP-I	14.06	12.11	11.53	18.24	17.9	11.22
CLIP-IA	35.3	30.21	29.32	39.93	39.27	28.86
CLIP-ITA	**45.85**	**41.04**	**40.02**	**50.04**	**49.87**	**39.69**

Conclusion

Summary

- Introduced category-to-image retrieval task.
- Introduced the model for the task.
- Evaluated the model in three settings: all categories, most general categories, most specific categories.
- Multimodal models tend to outperform unimodal models.
- Combining textual, visual, and attribute information when building product represetations produces best results on the task.

Summary

- Introduced category-to-image retrieval task.
- Introduced the model for the task.
- Evaluated the model in three settings: all categories, most general categories, most specific categories.
- Multimodal models tend to outperform unimodal models.
- Combining textual, visual, and attribute information when building product represetations produces best results on the task.

Thank you for your attention!

References i

[1] H. Bonab, M. Aliannejadi, A. Vardasbi, E. Kanoulas, and J. Allan. Cross-market product recommendation. In *CIKM*. ACM, 2021.

[2] Z. Dai, G. Lai, Y. Yang, and Q. V. Le. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *arXiv preprint arXiv:2006.03236*, 2020.

[3] L. Freedman. Building ecommerce content you can bank on, 2008.

[4] K. Goei, M. Hendriksen, and M. de Rijke. Tackling attribute fine-grainedness in cross-modal fashion search with multi-level features. In *SIGIR*. ACM, 2021.

[5] S. Hewawalpita and I. Perera. Multimodal user interaction framework for e-commerce. In *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pages 9–16. IEEE, 2019.

[6] L. B. Jabeur, L. Soulier, L. Tamine, and P. Mousset. A product feature-based user-centric ranking model for e-commerce search. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 174–186. Springer, 2016.

References ii

[7] N. Kondylidis, J. Zou, and E. Kanoulas. Category aware explainable conversational recommendation. *arXiv preprint arXiv:2103.08733*, 2021.

[8] K. Laenen and M.-F. Moens. Multimodal neural machine translation of fashion e-commerce descriptions. In *International Conference on Fashion communication: between tradition and future digital developments*, pages 46–57. Springer, 2019.

[9] K. Laenen, S. Zoghbi, and M.-F. Moens. Web search of fashion items with multimodal querying. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 342–350, 2018.

[10] H. Li, P. Yuan, S. Xu, Y. Wu, X. He, and B. Zhou. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8188–8195, 2020.

[11] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke. Improving outfit recommendation with co-supervision of fashion generation. In *The World Wide Web Conference*, pages 1095–1105, 2019.

References iii

[12] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[13] J. Tagliabue, B. Yu, and M. Beaulieu. How to grow a (product) tree: personalized category suggestions for ecommerce type-ahead. *arXiv preprint arXiv:2005.12781*, 2020.

[14] M. Tsagkias, T. H. King, S. Kallumadi, V. Murdock, and M. de Rijke. Challenges and research opportunities in ecommerce search and recommendations. *SIGIR Forum*, 54(1), June 2020.

[15] P. Wirojwatanakul and A. Wangperawong. Multi-label product categorization using multi-modal fusion models. *arXiv preprint arXiv:1907.00420*, 2019.

[16] T. Yashima, N. Okazaki, K. Inui, K. Yamaguchi, and T. Okatani. Learning to describe e-commerce images from noisy online data. In *Asian Conference on Computer Vision*, pages 85–100. Springer, 2016.

[17] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.