



UNIVERSITEIT VAN AMSTERDAM
Faculteit der Natuurwetenschappen,
Wiskunde en Informatica

Multimodal Machine Learning for Information Retrieval
Mariya Y. Hendriksen 2024

Multimodal Machine Learning for Information Retrieval

A Vision and Language Perspective

Mariya Y. Hendriksen

Multimodal Machine Learning for Information Retrieval

A Vision and Language Perspective

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties
ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 13 december 2024, te 14:00 uur

door

Mariya Yurievna Hendriksen

geboren te Goebkinski

PROMOTIECOMMISSIE

Promotor:

prof. dr. M. de Rijke

Universiteit van Amsterdam

Copromotor:

prof. dr. P.T. Groth

Universiteit van Amsterdam

Overige leden:

prof. dr. E. Kanoulas

Universiteit van Amsterdam

prof. dr. M.A. Larson

Radboud Universiteit

prof. dr. M.-F. Moens

Katholieke Universiteit Leuven

prof. dr. M. Worring

Universiteit van Amsterdam

dr. A.C. Yates

Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis has primarily been carried out at the Information Retrieval Lab of the University of Amsterdam and in part during an internship at Bloomberg AI. The research carried out at the University of Amsterdam was funded by Ahold Delhaize.

Copyright © 2024 by Mariya Yurievna Hendriksen, Amsterdam, The Netherlands.

Cover design adapted from “All Your Colours” by Simone Boon.

Printed by Ridderprint, The Netherlands.

ISBN: 978-94-6506-438-3

to my grandparents, Raqiyah and Yuriy.

ACKNOWLEDGEMENTS

Pursuing a PhD was never part of my original plan. I have always been interested in linguistics, mathematics, and computer science. This led me to a bachelor's in computational linguistics, where I discovered that it was the computational aspect of the field that fascinated me the most. From there, I pursued a master's in artificial intelligence, where I had an opportunity to engage in extracurricular research at ETH Zurich and KU Leuven. These experiences motivated me to commence a PhD journey – one that I am now close to completing.

The path I have taken has spanned six countries and helped me to grow both professionally and personally. What made the biggest difference were the wonderful people I met along the way. While I cannot mention everyone, I will do my best to acknowledge those who have had the strongest impact on my journey.

Maarten, thank you for your supervision. I have learned so much from you, and as I progress in my career and life, I gain an even greater appreciation of all you have taught me. I consider myself incredibly lucky to have had you as my key mentor and guide.

Paul, thank you for being my co-supervisor and for the engaging, insightful discussions. Your invaluable advice and cheerful attitude always kept me motivated.

Andrew, Evangelos, Marcel, Martha, and Sien, it is an honour to have you on my committee. Andrew, thank you for consistently offering valuable advice and thoughtful feedback on my ideas over the last few years. Evangelos, I am grateful for your leadership within the group and the insightful conversations we have shared. Martha, I appreciate you joining my committee, and I look forward to discussing my thesis with you. Marcel, thank you for the interesting conversations we had during my time at the University of Amsterdam and for dedicating your time to reading and discussing my thesis. Sien, you have been a role model for me since my master's, and it means a lot to have you on my committee.

During my time at AIRLab, I had the pleasure of collaborating with great people from Bol.com. Ernst, thank you for your guidance and for offering a broader perspective that shaped my research. Pim, Bart, Almer, Isaac, Karoliina, Binyam, Erick, and Joris – thank you for all your valuable contributions to this collaboration.

A huge thank you to Petra, Ivana, and Pablo for all your efforts in keeping the group running smoothly, and for the pleasant conversations. I truly appreciate all you have

done to make the group a better place for all of us.

My paranymphs, Maria and Harrie, thank you for standing by my side during the defence. Maria, thank you for sharing with me all a PhD life has to offer, for dinners, hikes, and, most of all, for your constant support and friendship. Harrie, on my very first day you were sitting next to me at the research meeting, talking about “wisdom” (WSDM). Thank you for sharing your wisdom with me ever since, and for a great time at conferences and parties.

Svitlana, your friendship and mentorship have had a profound impact on me, both as a researcher and as a person. I have learned so much from you. Сердечно дякую!

Mozhdeh, thank you for the great discussions and your support, the fun conferences and birthday parties, for all the jokes and all the hikes we have yet to do.

Sami, thank you for sitting across from me until the very end, for dinners and nights out followed by random walks around Madrid, and for everything else we have shared along the way.

Maartje, thank you for ice skating, naturalization ceremonies, and a great time in London – but especially for the insightful conversations, inspirational advice, and all your help.

I am particularly grateful to Andrew, Ana, Antonis, Shashank, Tom, Katya, Marzieh, Clara, Roxana, Julien, Romain, Gabriel, Thông, Spyretta, Jiahuan, Pooya, Ali, Sam, Christophe, Rolf, Artem, and Hosein. Each of you made this journey not just about research, but also about meaningful connections and shared memories.

I would also like to thank all the other members of ILPS/IRLab I met during my time at the University of Amsterdam: Alessio, Ali, Amin, Amir, Anna, Arezoo, Barrie, Bob, Carsten, Chang, Christof, Chuan, Chuan, Clemencia, Cosimo, Dan, David, David, Dylan, Evgenii, Federico, Gabrielle, Georgios, Hamid, Hinda, Hongyu, Ilias, Ilya, Iman, Ivana, Jasmin, Jie, Jin, Jingfen, Jingwei, Justine, Kidist, Maarten, Mahsa, Marta, Maurits, Ming, Mohammad, Mostafa, Mounia, Negin, Nikos, Olivier, Oscar, Pablo, Panagiotis, Peilei, Pengjie, Petra, Philipp, Ruben, Ruqing, Saedeh, Sebastian, Shaojie, Shubha, Simon, Songgaojun, Thilina, Trond, Vaishali, Vera, Wanyu, Weija, Xinyi, Yang, Yangjun, Yibin, Yifei, Yongkang, Yougang, Yuanna, Yuanxing, Yuyue, Zahra, Zihan, Ziyi, Zhaochun, and Ziming – thank you for all the fun and interesting research meetings, Soos and SEA talks, and vrijdagmiddagborrels.

Outside of my direct colleagues, I would like to thank Katrien, Nanne, Ivona, Phillip, Thomas, Jasmijn, David, and Tom for the inspiration and the insightful discussions.

To the students I supervised – Anne, Christina, Derek, Efstathios, Kenneth, Luka, Maurice, and Viggo – thank you for teaching me more than I ever taught you.

During my studies, I had the opportunity to undertake several internships, each of which was an invaluable experience: • *ETH Zurich*: It was here that I decided to pursue a PhD. Gunnar and Vincent, thank you for your leadership and mentorship.

Megha, Evgenii, and Thomas, I cherish the memories from E 62.1-62.2 and the experiences we built together. I would also like to thank Rana, Joshua, Ciara, Cosmina, Ljubica, Pantea, Elisabeth, Denys, Mihaela, and Polina for lunches at 12:27:35, hikes, parties, and all the moments we shared. • *LIIR (KU Leuven)*: Sien, your leadership and insightful discussions were invaluable to me, and Tuur, thank you for your mentorship. Katrien, I am grateful for our discussions about multimodality. My gratitude extends to Thierry, Shurong, Guillem, Graham, and Daniel. • *Amazon Science*: Yunlong, thank you for your friendship and mentorship since my last internship week and up until now. Danae, Francesco, Zhengxiang, and Yue – thank you for a great time and engaging discussions. I would also like to thank Gabriella, Nikos, Darragh, Jordan, and Emine. • *Bloomberg AI*: Shuo, I learned so much from you – thank you for your mentorship. Mohamed, thank you for your leadership in the team. I am particularly grateful to Ridho, Edgar, Diego, Chengshuai, Jeff, Sean, Guillaume, Steven, Anju, Bruno, Francesca, Halil, Irinka, Lewis, Marco, Rumana, and Sawan. • *Google*: Aleksandr, thank you for hosting me, and Aliaksei, I appreciate your leadership in the team. I am especially grateful to Sertan, Andreas, Lucas, Bilal, François, Nastia, and Alexander, from whom I learned a lot. I would also like to thank Andrea, Polina, Jakub, Jonathan, Vikas, Alizée, Robert, Matej, Daniil, Chintu, Sebastian, Taylan, Arthur, Daniele, Dmytro, Lam, Noah, Igor, Tomasz, Khalid, Jan, Doron, Noam, and Yonatan. • *Microsoft Research*: I am especially thankful to Sam for hosting me and Katja for her leadership, insightful discussions, and guidance. I am grateful to Chentian, Raluca, Dave, Abdelhak, Sarah, Tabish, Teodora, Yuhan, Sergio, Shanzheng, Linda, Max, Liana, Nando, and Tadas.

To the friends I have made outside of the PhD life along the way – Katya, Annebeth, Yunlong, Anton, Khumar, Polina, Nathalie, Dasha, and Prune – thank you for being a part of my life, even from afar.

I am grateful to Dini, Wilbert, Birgitt, Art, Freke, Gerrit, Annemiek, Wim, Elly, Erik as well as Bart, Bart, Bart, Arjan, Corstian, Daniëlle, Marco, Mathijs, Ida, Maarten, Maurice, Stijn, Helen, and Michelle for enriching my life in the Netherlands. Stijn, a special thanks to you for keeping my plants alive while I was away! Angela, thank you for your support over the past few years. Rinke, thank you for all the great adventures since 2014, for bringing fresh perspectives into my life, and for your support.

I would like to conclude by thanking my friends and family from Russia and Ukraine, whom I wish I could see more often. Valeriy Sergeevich, thank you so much for everything you have taught me – our conversations about life in English have significantly influenced my worldview. Silviya, Guzel, and Nastia – thank you for being part of my life since I was four, ten, and twelve. Your friendship has had a lasting impact on me.

I am grateful to my aunts, uncles, and cousins: Yulia, Elia, Polina, Luiza, Radik,

Kristina, Alina, Sasha, and Galya – thank you for your support. A special thanks to my grandparents: Varvara and Alexander, although I only knew you from my father's stories, I am sincerely grateful for everything you did. Raqiyah and Yuriy – thank you for our trips to the Sea of Azov and the garden, for your profound wisdom, and for everything you taught me. I am deeply sorry that you cannot be present at my defence.

My brothers, Misha and Lesha, thank you for your friendship, support, and understanding; for knowing me so well, despite the distances between us. My parents, Viktoriya and Yuriy: thank you for everything you have done for me, for encouraging my curiosity, for setting an example of independence and confidence, and for showing me how to face whatever comes in life with a dose of humour. The latter has proven to be especially invaluable on this journey.

Я хочу поблагодарить своих друзей и родных из России и Украины, которых мне хотелось бы видеть чаще. Валерий Сергеевич, большое спасибо за все, чему Вы меня научили. Наши беседы о жизни на английском значительно повлияли на мое восприятие мира. Сильвия, Гузель и Настя – спасибо, что вы были в моей жизни с четырёх, десяти и двенадцати лет. Ваша дружба оказала большое влияние на меня.

Также хочу выразить свою благодарность моим тетям, дядям, и кузенам: Юля, Эля, Полина, Луиза, Радик, Кристина, Алина, Саша и Галя – спасибо за вашу поддержку. Отдельное спасибо моим бабушкам и дедушкам: Варвара и Александр, хотя я знала вас только по рассказам отца, я искренне благодарна за все, что вы сделали. Ракия и Юрий – спасибо за наши путешествия на Азовское море и огород, за вашу безграничную мудрость и за все, чему вы меня научили. Мне очень жаль, что вы не можете присутствовать на моей защите.

Мои братья, Миша и Леша, спасибо за дружбу, поддержку и понимание; за то, что вы так хорошо меня знаете, несмотря на расстояния. Мои родители, Виктория и Юрий: спасибо за все, что вы для меня сделали, за поощрение моего любопытства, за пример независимости и уверенности в себе, а также за то, что показали, как относится ко всему в жизни с долей юмора. Последнее мне особенно пригодилось в пути.

Mariya Hendriksen
London and Cambridge
Autumn of 2024.

CONTENTS

Acknowledgements	v
1 Introduction	1
1.1 Research Outline and Questions	3
1.2 Main Contributions	6
1.3 Thesis Overview	9
1.4 Origins	9
2 Scene-Centric vs. Object-Centric Image-Text Retrieval	13
2.1 Introduction	14
2.2 Related Work	16
2.2.1 Cross-Modal Retrieval	16
2.2.2 Scene-Centric and Object-Centric Datasets	17
2.2.3 Reproducibility in Cross-Modal Retrieval	17
2.3 Task Definition	18
2.4 Methods	18
2.4.1 Methods Included for Comparison	18
2.4.2 Methods Excluded from Comparison	19
2.5 Experimental Setup	20
2.5.1 Datasets	20
2.5.2 Subtasks	21
2.5.3 Experiments	21
2.6 Results	23
2.6.1 RQ1.1: Reproducibility	23
2.6.2 RQ1.2: Replicability	25
2.6.3 RQ1.3: Generalizability	26
2.7 Discussion and Conclusion	27
3 Multimodal Learned Sparse Retrieval	29
3.1 Introduction	30
3.2 Related Work	31
3.2.1 Learned Sparse Retrieval	31
3.2.2 Cross-Modal Retrieval	32
3.3 Background	33

3.4	Methodology	34
3.4.1	Model Architecture	34
3.4.2	Training Loss	37
3.5	Experiments and Results	37
3.5.1	Experimental Setup	37
3.5.2	Results and Discussion	38
3.5.3	Retrieval Latency of Dense and Sparsified Models	43
3.6	Conclusion	44
4	Shortcuts in Vision-Language Representation Learning	45
4.1	Introduction	46
4.2	Background and Analysis	50
4.2.1	Preliminaries	50
4.2.2	Analysis of Contrastive Vision-Language Representation Learning for Multiple Captions per Image	51
4.3	Synthetic Shortcuts to Control Shared Information	54
4.4	Synthetic Shortcuts and their Impact on Learned Representations	56
4.4.1	Findings	57
4.4.2	Upshot	58
4.5	Reducing Shortcut Learning	59
4.5.1	Latent Target Decoding	59
4.5.2	Implicit Feature Modification	60
4.5.3	Method Comparison	60
4.6	Experimental Results	61
4.6.1	Does Latent Target Decoding Reduce Shortcut Learning?	61
4.6.2	Does Implicit Feature Modification Reduce Shortcut Learning?	61
4.6.3	Upshot	64
4.7	Related work	64
4.7.1	Multi-View Representation Learning	64
4.7.2	Vision-language Representation Learning	65
4.7.3	Our Focus	68
4.8	Conclusion	68
4.9	Broader Impact	70
	Appendices	71
4.A	Notation	71
4.B	Problem Definition and Assumptions	72
4.B.1	Evaluation Task	72
4.B.2	Assumptions	72
4.C	Analysis of Contrastive Learning for Multiple Captions per Image	73

4.D	Experimental Setup	74
4.D.1	Datasets	74
4.D.2	Models	75
4.D.3	Training	75
4.D.4	Shortcut Sampling	76
4.E	Optimization Objectives	76
4.E.1	InfoNCE	76
4.E.2	Latent Target Decoding	76
4.E.3	Implicit Feature Modification	78
5	Assessing Brittleness of Image-Text Retrieval Benchmarks	79
5.1	Introduction	80
5.2	Preliminaries	82
5.3	Concept Granularity in Image-Text Retrieval Datasets	83
5.3.1	Granularity Features in Image-Text Retrieval	83
5.3.2	Granularity Analysis	84
5.4	Evaluation Framework	85
5.4.1	Perturbations	85
5.4.2	Evaluation Metric	88
5.5	Experiments	89
5.5.1	Models	89
5.5.2	Experiments Overview	90
5.5.3	Results	91
5.5.4	Model Input Analysis	94
5.6	Related Work	94
5.6.1	Cross-Modal Retrieval	94
5.6.2	Vision-Language Model Evaluation	96
5.7	Conclusion	97
6	Predicting Purchase Intent for Product Retrieval	99
6.1	Introduction	100
6.2	Background and Definitions	102
6.3	Dataset Description	103
6.4	Characterizing Purchase Intent	104
6.4.1	Session Length	105
6.4.2	Temporal Variations	106
6.4.3	Channel Types	107
6.4.4	Devices	109
6.4.5	Queries	112
6.4.6	Purchase Intent Characteristics	114

6.5	Predicting Purchase Intent	114
6.5.1	Experimental Setup	115
6.5.2	Prediction for Anonymous Users	116
6.5.3	Prediction of Identified Users	117
6.5.4	Feature Importance Analysis	117
6.6	Related Work	119
6.6.1	E-Commerce User Purchase Behavior Analysis	119
6.6.2	Purchase Prediction in E-Commerce	120
6.7	Discussion and Conclusion	121
7	Extending CLIP for Category-to-Image Retrieval	123
7.1	Introduction	124
7.2	Related Work	125
7.2.1	Learning Multimodal Embeddings.	125
7.2.2	Multimodal Image Retrieval	126
7.2.3	Multimodal Retrieval in E-Commerce	126
7.3	Approach	127
7.3.1	Task Definition	127
7.3.2	CLIP-ITA	127
7.4	Experimental Setup	129
7.4.1	Dataset.	129
7.4.2	Evaluation Method	130
7.5	Experimental Results	131
7.5.1	Baselines.	131
7.5.2	Image-Based Representations	133
7.5.3	Image- and Attribute-Based Representations	133
7.5.4	Image-, Attribute-, and Title-Based Representations	133
7.6	Error Analysis	134
7.6.1	Distance between Predicted and Target Categories.	134
7.6.2	Performance on Seen vs. Unseen Categories	135
7.7	Conclusion	136
8	Conclusion	139
8.1	Summary of Findings	140
8.2	Future Work	144
8.3	Final Remarks	145
	Bibliography	147
	Summary	166

Samenvatting

168

1

INTRODUCTION

Suppose you want to learn more about a certain topic. For the sake of argument, let us assume this topic is multimodal machine learning for information retrieval. How would you approach the task? You could continue to read this thesis, attend relevant lectures, or talk to experts in the field. Essentially, you would seek, gather, and process relevant information from diverse sensory inputs or modalities, like visual and auditory, to develop an understanding of the subject.

This would be possible because our brains are designed to process information from different sensory inputs in interconnected regions (Binder et al., 2009). A canonical example of this is the McGurk effect (McGurk and MacDonald, 1976), a perceptual phenomenon demonstrating how combining of visual and auditory signals influences speech perception. This cross-modal connectivity extends to how the brain links visual and linguistic processing. Visual information, processed in the occipital cortex (Grill-Spector et al., 2001), is closely connected to Wernicke's area, which is important for understanding language (Binder, 2015). Studies have shown that our brain integrates these inputs, enabling us to simultaneously recognize objects and understand their descriptions (Tomasello et al., 2017).

Inspired by the human ability to learn by processing multimodal sensory inputs, the goal of multimodal machine learning is to create models that can handle and relate information from multiple modalities (Baltrusaitis et al., 2019). This research domain brings some unique challenges due to the heterogeneity gap between different modalities (Carvalho et al., 2018; Hu et al., 2019) and the complementary and redundant nature of multimodal data (Baltrusaitis et al., 2019).

A key aspect of learning is discovery – the process of seeking and finding relevant information. Humans excel in this task by leveraging their ability to integrate information from multiple sensory inputs. Similarly, multimodal machine learning for information retrieval aims to replicate this ability by enabling models to process and relate data from different modalities (Baltrusaitis et al., 2019; Laenen, 2022). This

approach allows retrieval systems to provide richer, more accurate retrieval results, mirroring how humans form a clearer understanding by combining various sensory inputs.

Motivated by this topic, this thesis focuses on multimodal machine learning for information retrieval as the main task, specifically from a vision and language standpoint. We are particularly interested in image-text retrieval, a bidirectional retrieval task across image and text data. We explore three key areas within this domain:

- I *Dense and Sparse Retrieval*: We focus on retrieving information across image and text data in dense and sparse retrieval settings. We start our investigation by examining foundation vision-language models and their reproducibility, replicability, and generalizability in the context of the image-text task on both scene-centric and object-centric datasets. Here, a dataset is called scene-centric if it depicts complex scenes with multiple objects and their interactions, and object-centric if features single objects with detailed descriptions (Zhang et al., 2021a; Shen et al., 2019). We then explore how vision-language models can be adapted for learned sparse retrieval, addressing the challenges of sparsification in the vision-language domain. We define vision-language domain as research that integrates and aligns information from image data and textual data to facilitate image-text retrieval (Baltrusaitis et al., 2019).
- II *Representation Learning and Evaluation*: We investigate the quality of learned multimodal representation and evaluation procedures. We consider the problem of learning shortcuts, i.e., easy-to-detect discriminatory features that minimize optimization objectives but do not represent all the information needed for solving the task at hand (Geirhos et al., 2020; Hermann and Lampinen, 2020; Robinson et al., 2021). We focus on the challenge of shortcut learning in contrastive learning with multiple captions per image and propose a framework for controlled investigation of this problem. We continue our investigation by exploring the brittleness of existing image-text retrieval evaluation pipeline. By brittleness of an evaluation pipeline we mean the vulnerability of vision-language models to performance degradation when faced with more complex or varied inputs than those found in standard benchmarks (Chen et al., 2023b). We are especially interested in brittleness as it relates to concept granularity, by which we mean the level of detail or specificity in the relationship between images and their corresponding textual descriptions (Laenen et al., 2018; Pesahov et al., 2023). During our investigation we focus on both existing datasets and evaluation metrics. We propose an evaluation suite to address the highlighted problems.
- III *Product Retrieval*: We explore the application of multimodal machine learning in product retrieval. Starting with an analysis of search logs from a European

e-commerce platform, we examine sessions across different devices, each with a unique set of modalities. We focus on modelling purchase intent and exploring how different modalities available on a given device impact user behaviour. Motivated by our findings, we propose the category-to-image retrieval task, relevant to e-commerce applications, and develop a model for the task. We investigate how multimodal representations affect model performance on the task across categories of varying granularity.

By investigating these core areas, we aim to contribute to the field of multimodal machine learning for information retrieval through novel algorithmic, empirical, resource, and theoretical contributions.

1.1 RESEARCH OUTLINE AND QUESTIONS

We center our research around six research questions, each of which we address in a dedicated chapter of this thesis. Below, we provide an overview of the questions.

We start our investigation in the domain of dense and sparse retrieval. We examine the reproducibility of image-text cross-modal retrieval results across scene-centric and object-centric datasets. As we explained earlier in this chapter, scene-centric datasets depict complex scenes with multiple objects and their interactions, while object-centric datasets feature single objects with detailed descriptions (Zhang et al., 2021a; Shen et al., 2019). Previous work on image-text cross-modal retrieval has predominantly focused on scene-centric benchmarks, leaving object-centric datasets relatively under-explored (Zhou et al., 2014).

Motivated by this gap, we focus on the reproducibility, replicability, and generalizability of published relative performance image-text cross-modal retrieval results across scene-centric and object-centric datasets.

Therefore, we formulate our first research question as follows:

RQ1 To what extent are the published image-text cross-modal retrieval results reproducible, replicable, and generalizable across scene-centric and object-centric datasets?

To address this question, we conduct a reproducibility study using two state-of-the-art cross-modal retrieval models (Zeng et al., 2022; Radford et al., 2021). We evaluate these models on two scene-centric datasets (Lin et al., 2014; Young et al., 2014) and three object-centric datasets (Welinder et al., 2010; Collins et al., 2022; Han et al., 2017). Our results show that while relative performance results are partially reproducible on scene-centric datasets, they face challenges on object-centric datasets. Besides, the

absolute performance scores on object-centric datasets are lower compared to scene-centric datasets. This work highlights the importance of exploration and evaluation of cross-modal retrieval methods across diverse benchmarks and contributes to our understanding of the capabilities of cross-modal retrieval models as well as areas for improvement.

We continue investigating dense and sparse retrieval from the perspective of multimodal learned sparse retrieval. Learned sparse retrieval approaches encode queries and documents into sparse lexical vectors, offering potential interpretability and leveraging traditional inverted index structures (Formal et al., 2021; Formal et al., 2022; Nguyen et al., 2023b). However, little is known about their generalizability in the vision-language domain. As we explained earlier in this chapter, we define the vision-language domain as research that integrates and aligns information from image data and textual data to facilitate image-text retrieval (Baltrusaitis et al., 2019). Motivated by this gap, we ask the following research question:

RQ2 How can learned sparse retrieval techniques be applied in the vision-language domain?

To answer this question, we design a method for multimodal learned sparse retrieval and investigate its performance on the image-text retrieval task. During the evaluation, we discover the phenomena of dimension co-activation and semantic deviation, and propose metrics to quantify them. We further evaluate the model performance using both already defined and additional metrics. Our findings show that the proposed method effectively converts dense to sparse representations, and maintains competitiveness with dense models. We demonstrate how the discovered phenomena can be partially mitigated using query expansion control during training.

We continue our investigation in the space of representation learning and evaluation. We start by considering the problem of shortcut learning in the context of vision-language contrastive learning with multiple captions per image. As we explained earlier in this chapter, we define shortcuts as easy-to-detect discriminatory features that minimize optimization objectives but do not represent all the information needed for solving the task at hand (Geirhos et al., 2020; Hermann and Lampinen, 2020; Robinson et al., 2021). We focus on the situation when all captions associated with an image contain both shared and caption-specific information. We investigate if it is possible to contrastively learn task-optimal vision-language representations in this context, and formulate our research question as follows:

RQ3 In the context of vision-language representation learning with multiple captions per image, to what extent does the presence of a shortcut hinder learning task-optimal representations?

To address this question, we introduce a synthetic shortcuts for vision-language framework, a novel framework which augments image-caption tuples with identifiers that do not bear any semantic meaning. We use this framework to analyze the extent to which a vision-language model relies on synthetic shortcuts during training and evaluation and explore how shortcut learning can be mitigated. We show that contrastive vision-language methods tend to depend on shortcuts and suppress task-relevant information in the context of multiple captions per image.

Next, we address the brittleness of evaluation and benchmarking of vision-language models on the image-text retrieval task, with a focus on concept granularity (Laenen, 2022; Zhao et al., 2022). As we explained earlier in this chapter, brittleness refers to the vulnerability of vision-language models to performance degradation when faced with more complex or varied inputs than those found in standard benchmarks (Chen et al., 2023b). Concept granularity refers to the level of detail or specificity in the relationship between images and their corresponding textual descriptions (Laenen et al., 2018; Pesahov et al., 2023). Current benchmarks often lack the necessary level of detail in textual descriptions, leading to coarse-grained datasets that may not fully capture the relationships between images and text (Chen et al., 2023b; Goei et al., 2021). Motivated by this problem, we ask the following research question:

RQ4 How can we improve the evaluation and benchmarking of vision-language models on the image-text retrieval task?

To answer this research question, we analyze concept granularity within existing image-text retrieval benchmarks, comparing them with fine-grained counterparts. We propose a novel evaluation framework incorporating perturbations and a new metric to capture semantic similarity and cross-modal relationships. We evaluate four state-of-the-art vision-language models on this framework, assessing reproducibility and sensitivity to perturbations on both coarse and fine-grained datasets. This work contributes to our understanding of the impact of concept granularity on model performance on the image-text retrieval task and opens up potential directions for refining evaluation and benchmarking processes for the task.

We shift our investigation to product retrieval and focus on the problem of purchase intent prediction in a cross-device scenario, where each device represents information given a unique set of modalities (Montanez et al., 2014). We consider the problem in the context of anonymous vs. identified sessions. The majority of work on the topic of predicting purchase intent for product retrieval has focused on known customers, ignoring anonymous sessions (Tsagkias et al., 2020). Therefore, motivated by this gap, we aim to answer the following research question:

RQ5 How can we facilitate product retrieval by predicting purchase intent in cross-device setting?

To answer this research question, we sample and analyze session logs from a European e-commerce platform and identify purchase intent signals like session duration, timing, device type, channel, and queries. We design features based on these insights, develop predictive models for both session types, and evaluate model performance. This work contributes to our understanding of user behaviour across devices.

Motivated by the findings from the previous chapter, for our final research question, we propose and motivate the category-to-image retrieval task. The goal of the task is to retrieve relevant images of items associated with a given category sampled from a category tree. The category-to-image retrieval task is important in information retrieval, particularly in the context of retrieving images of concepts of varying granularity (Kondylidis et al., 2021). Users often encounter challenges in matching textual descriptions with corresponding visual representations across a spectrum of categories (Tagliabue et al., 2020; Nielsen et al., 2000). This mismatch can lead to sub-optimal search and recommendation results, affecting overall system performance. To address this issue, we focus on exploring how building multimodal product representations can impact the performance on the task. This motivates the research question:

RQ6 How do multimodal document representation, encompassing text, image, and attribute data, impact the performance on the category-to-image retrieval in the context of categories of varying granularity?

To answer this research question, we conduct a series of experiments. First, we adapt an e-commerce dataset containing textual descriptions, images, and attribute information of products across a diverse set of categories and prepare a set of unimodal and bi-modal baselines for the task. Next, we design and implement a multimodal retrieval model, which combines textual, visual, and attribute information to create product representations. We show that multimodal document representation generally improves performance on the task. Notably, on the most general categories, models incorporating image information alone demonstrate slightly better performance, highlighting the relevance of visual cues in identifying broader product categories. This work helps us to understand how different modalities impact the models' performance on the task in the context of concepts of varying granularity.

This concludes the overview of our research questions. In the next section, we summarize the main contributions of this thesis.

1.2 MAIN CONTRIBUTIONS

In this section, we summarize the main contributions of this thesis. We divide the contributions in this thesis into algorithmic, empirical, resource, and theoretical con-

tributions. Together, these contributions help us gain deeper insights into cross-modal retrieval, contrastive vision-language representation learning, and product retrieval.

Algorithmic contributions

- We train a lightweight projection head to convert dense to sparse vectors for multimodal learned sparse retrieval. We show that our sparsified models are faithful to dense models while delivering competitive results (Chapter 3).
- We propose the framework for synthetic shortcuts in vision-language models, a novel training and evaluation framework that allows us to inject synthetic shortcuts into image-text data to measure to what extent contrastive image-text methods rely on shortcuts to minimize the contrastive optimization objective (Chapter 4).
- We present two shortcut learning reduction methods on our proposed training and evaluation framework (Chapter 4).
- We propose a novel framework for evaluating vision-language models on the image-text retrieval task, which includes word-level and caption-level perturbations for model inputs and a cross-modal evaluation metric (Chapter 5).
- We propose CLIP-ITA, the first model specifically designed for the category-to-image retrieval task. CLIP-ITA leverages multimodal product data such as textual, visual, and attribute data (Chapter 7).

Empirical contributions

- We investigate reproducibility, replicability and generalizability of cross-modal retrieval results for scene-centric and object-centric datasets (Chapter 2).
- We evaluate the framework for synthetic shortcuts in vision-language models on a variety of settings and demonstrate that increasing the number of shortcuts in the training data induces contrastive image-text methods to rely on these shortcuts, leading to a suppression of task-relevant information (Chapter 4).
- We examine two shortcut reduction methods (latent target decoding and implicit feature modification) on the framework for synthetic shortcuts in vision-language models and show that these methods can partially mitigate the shortcut learning problem in some settings (Chapter 4).
- We evaluate the impact of dataset granularity on the performance of vision-

language models on the image-text retrieval task using standard benchmarks, MS-COCO and Flickr-30k, and their fine-grained counterparts (Chapter 5).

- We conduct a comprehensive evaluation of four state-of-the-art vision-language models using our proposed framework. This includes examining the reproducibility of scores, the impact of perturbations on zero-shot performance on the image-text retrieval task, and model behaviour with respect to dataset granularity (Chapter 5).
- We define two feature sets for modelling purchase for product retrieval, tailored towards anonymous sessions and identified sessions (Chapter 6).
- We evaluate our proposed features by extending an existing production-ready model and running additional experiments with classifiers generally used for this task (Chapter 6).

Resource contributions

- We conduct an in-depth analysis of a real-world customer interaction dataset with more than 95 million sessions, sampled from a large European e-commerce platform. We identify session features such as device type and conversion rate, weekday, channel type, and features based on historic customer data to distinguish between purchase and non-purchase sessions (Chapter 6).

Theoretical contributions

- We propose a line of research for efficiently converting a multi-modal dense retrieval model to a multi-modal learned sparse retrieval model (Chapter 3).
- We identify the issues of high dimension co-activation and semantic deviation and propose a training method to address them (Chapter 3).
- We prove that contrastive losses that enforce minimal sufficient representations can never learn task-optimal image representations (i.e., representations that contain all task-relevant information in the input captions), in the context of image-text representation learning with multiple matching captions per image (Chapter 4).
- We propose and motivate the task of category-to-image retrieval, a novel task of retrieving an image given a category of varying granularity (Chapter 7).

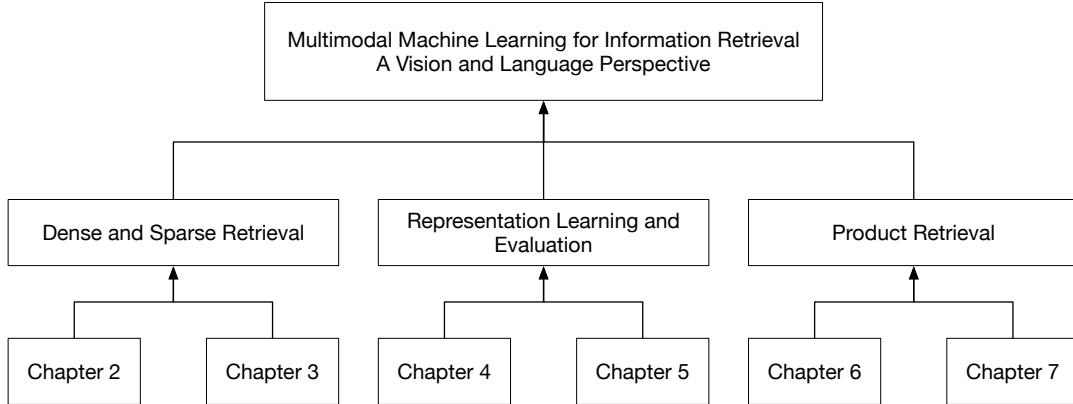


Figure 1.1: Thesis outline w.r.t. its key areas.

1.3 THESIS OVERVIEW

In this section, we present an overview of the thesis and offer guidance on how to navigate through its contents. The thesis consists of eight chapters, with the current chapter being the first.

The following six chapters explore core research questions outlined in Section 1.1 and focus on three main themes: dense and sparse retrieval, representation learning and evaluation, and product retrieval. Figure 1.1 illustrates the chapter structure mapped to these themes. Each chapter builds upon a single research paper (detailed in Section 1.4), and can be read independently. Finally, Chapter 8 summarizes the thesis findings and explores potential future research directions.

1.4 ORIGINS

Below we list the publications that are the origins of each chapter.

Chapter 2 is based on the following paper:

- **Mariya Hendriksen**, Svitlana Vakulenko, Ernst Kuipers, and Maarten de Rijke. Scene-Centric vs. Object-Centric Image-Text Cross-Modal Retrieval: A Reproducibility Study. In *ECIR 2023: 45th European Conference on Information Retrieval* (Hendriksen et al., 2023).

MH: Conceptualization, Methodology, Investigation, Software, Writing – Original Draft, Writing – Review & Editing, Project Administration. SV: Supervision, Methodology, Writing – Review & Editing. EK: Supervision, Resources. MdR: Funding Acquisition, Supervision, Methodology, Writing – Review & Editing.

Chapter 3 is based on the following paper:

- Thong Nguyen, **Mariya Hendriksen**, Andrew Yates, and Maarten de Rijke.

Multimodal Learned Sparse Retrieval with Probabilistic Query Expansion.
In *ECIR 2024: 46th European Conference on Information Retrieval* (Nguyen et al., 2024).

MH and TN shared first authorship. TN: Conceptualization, Methodology, Investigation, Software, Writing – Original Draft, Writing – Review & Editing. MH: Conceptualization, Methodology, Investigation, Writing – Original Draft, Writing – Review & Editing, Project Administration. AY, MdR: Funding Acquisition, Supervision, Methodology, Writing – Review & Editing.

Chapter 4 is based on the following paper:

- Maurits Bleeker, **Mariya Hendriksen**, Andrew Yates, and Maarten de Rijke. Demonstrating and Reducing Shortcuts in Vision-Language Representation Learning. In *TMLR: Transactions on Machine Learning Research* (Bleeker et al., 2024).

MH and MB shared first authorship. MB: Conceptualization, Methodology, Investigation, Software, Writing – Original Draft, Writing – Review & Editing. MH: Methodology, Formal Analysis, Investigation, Software, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration. AY: Supervision, Investigation, Writing – Review & Editing. MdR: Funding Acquisition, Supervision, Methodology, Writing – Review & Editing.

Chapter 5 is based on the following paper:

- **Mariya Hendriksen**, Shuo Zhang, Ridho Reinanda, Mohamed Yahya, Edgar Meij, and Maarten de Rijke. Assessing Brittleness of Image-Text Retrieval Benchmarks from Vision-Language Models Perspective. *Under Submission* (Hendriksen et al., 2024).

MH: Conceptualization, Methodology, Investigation, Software, Writing – Original Draft, Writing – Review & Editing, Project Administration. SZ: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Project Administration. RR, MY, EM: Supervision, Investigation, Writing – Review & Editing. MdR: Supervision, Writing – Review & Editing. This work was done during an internship at Bloomberg AI in 2023.

Chapter 6 is based on the following paper:

- **Mariya Hendriksen**, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke. Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers. In *Proceedings of the 2020 SIGIR Workshop on eCommerce* (Hendriksen et al., 2020).

MH: Conceptualization, Methodology, Investigation, Software, Writing – Original Draft, Writing – Review & Editing, Project Administration. EK, PN, SS: Supervision, Methodology, Writing – Review & Editing. MdR: Funding Acquisition, Supervision, Methodology, Writing – Review & Editing.

Chapter 7 is based on the following paper:

- **Mariya Hendriksen**, Maurits Bleeker, Svitlana Vakulenko, Nanne van Noord, Ernst Kuipers, and Maarten de Rijke. Extending CLIP for Category-to-image Retrieval in E-commerce. In *ECIR 2022: 44th European Conference on Information Retrieval* (Hendriksen et al., 2022).

MH: Conceptualization, Methodology, Investigation, Software, Writing – Original Draft, Writing – Review & Editing, Project Administration. MB: Writing – Review & Editing. SV, NvN, EK: Supervision, Methodology, Writing – Review & Editing. MdR: Funding Acquisition, Supervision, Methodology, Writing – Review & Editing.

The writing of this thesis also benefited from work on the following publications:

- **Mariya Hendriksen**, Artuur Leeuwenberg, and Marie-Francine Moens. LSTM for Dialogue Breakdown Detection: Exploration of Different Model Types and Word Embeddings. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction* (Hendriksen et al., 2021).
- Goei, Kenneth, **Mariya Hendriksen**, and Maarten de Rijke. Tackling Attribute Fine-grainedness in Cross-modal Fashion Search with Multi-level Features. In *Proceedings of the 2021 SIGIR Workshop on eCommerce* (Goei et al., 2021).
- **Mariya Hendriksen**. Multimodal Retrieval in E-commerce: from Categories to Images, Text, and Back. In *ECIR 2022: 44th European Conference on Information Retrieval* (Hendriksen, 2022).
- **Mariya Hendriksen**, Viggo Overes, Luka Wong Chung. Unimodal vs. Multi-modal Siamese Networks for Outfit Completion (Hendriksen and Overes, 2022).
- Karel Veldkamp, **Mariya Hendriksen**, Zoltán Szlávik, and Alexander Keijser. Towards Contrastive Learning in Music Video Domain (Veldkamp et al., 2023).
- Thong Nguyen, **Mariya Hendriksen**, and Andrew Yates. Multimodal Learned Sparse Retrieval for Image Suggestion. In: *TREC 2023* (Nguyen et al., 2023a).

2

SCENE-CENTRIC VS. OBJECT-CENTRIC IMAGE-TEXT RETRIEVAL

We start our investigation by considering the reproducibility of image-text cross-modal retrieval (CMR) results. Reproducibility is important for understanding the robustness and effectiveness of state-of-the-art (SOTA) CMR methods. In this chapter, we focus on reproducibility, replicability, and generalizability of CMR results on scene-centric and object-centric datasets. *Scene-centric* datasets are collections of image-text pairs where each image depicts complex scenes containing multiple objects and their interactions. The corresponding text descriptions typically focus on conveying the entire scene, including the relationships and activities among the depicted objects. *Object-centric* datasets, on the other hand, consist of image-text pairs where each image features a single object of interest. These objects are often positioned centrally within the image. The corresponding text descriptions typically describe the depicted object and its fine-grained attributes. While the majority of work in CMR is conducted using scene-centric benchmarks, the performance of models on object-centric datasets remains relatively underexplored. Motivated by this gap, we ask the following research question:

RQ1: To what extent are the published image-text cross-modal retrieval results reproducible, replicable, and generalizable across scene-centric and object-centric datasets?

To answer this RQ, we select two models, CLIP and X-VLM, both models were considered SOTA on the CMR task at the moment of publication. Next, we evaluate the selected models on two scene-centric datasets (MS COCO, Flickr30k) and three object-centric datasets (CUB-200, Fashion200k, and ABO). We show that while relative per-

This chapter was published at the 45th European Conference on Information Retrieval (ECIR 2023) under the title “Scene-Centric vs. Object-Centric Image-Text Cross-Modal Retrieval: A Reproducibility Study” (Hendriksen et al., 2023).

formance results in cross-modal retrieval are partially reproducible on scene-centric ones when it comes to replicating the results on object-centric datasets, the relative performance results are not reproducible. Besides, the absolute performance scores on object-centric datasets are lower compared to scene-centric datasets. Hence, this chapter emphasises the importance of further exploration and evaluation of CMR methods across diverse benchmarks and contributes to our understanding of the capabilities of CMR models and areas for improvement.

2.1 INTRODUCTION

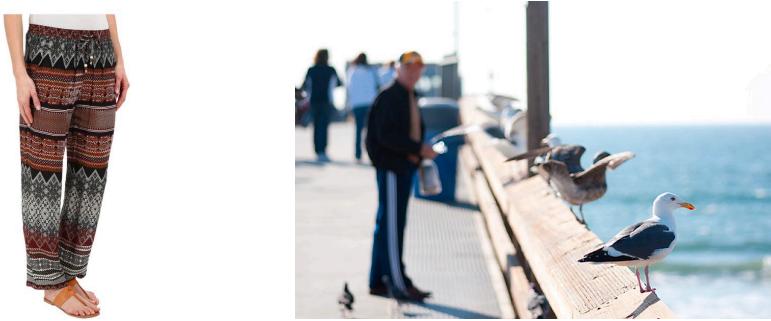
CMR is the task of finding relevant items across different modalities. For example, given an image, find a text or vice versa. The main challenge in CMR is known as *the heterogeneity gap* (Carvalho et al., 2018; Hu et al., 2019). Since items from different modalities have different data types, the similarity between them cannot be measured directly. Therefore, the majority of CMR methods published to date attempt to bridge this gap by learning a latent representation space, where the similarity between items from different modalities can be measured (Wang et al., 2016a).

In this work, we specifically focus on *image-text* CMR, which uses textual and visual data. The retrieval task is performed on *image-text pairs*. In each image-text pair, the text (often referred to as *caption*) describes the corresponding image it is aligned with. For image-text CMR we use either an image or a text as a query (Wang et al., 2016a). Hence, the CMR task that we address in this chapter consists of two subtasks: (i) *text-to-image retrieval*: given a text that describes an image, retrieve all the images that match this description; and (ii) *image-to-text retrieval*: given an image, retrieve all texts that can be used to describe this image.

Scene-centric vs. object-centric datasets. Existing image datasets can be grouped into *scene-centric* and *object-centric* datasets (Zhang et al., 2021a; Shen et al., 2019). The two types of datasets are typically used for different tasks, viz. the tasks of scene and object understanding, respectively. They differ in important ways that are of interest to us when evaluating performance and generalization abilities of CMR models.

Scene-centric images depict complex scenes that typically feature multiple objects and relations between them. These datasets contain image-text pairs, where, in each pair, an image depicts a complex scene of objects and the corresponding text describes the whole scene, often focusing on *relations and activities*.

Images in object-centric image datasets are usually focused on a single object of interest that they primarily depict. This object is often positioned close to the center of an image with other objects, optionally, in the background. Object-centric datasets contain image-text pairs, where, in each pair, an image depicts an object of interest



Multicolor boho batic pants

Seagulls sitting on the ledge of a pier
with people watching

Figure 2.1: An object-centric (left) and a scene-centric (right) image-text pair. Sources: Fashionzook (left); MS COCO (right).

and the corresponding text describes the depicted *object and its (fine-grained) attributes*.

To illustrate the differences between the two dataset types in CMR, we consider the examples provided in Figure 2.1 with an object-centric image-caption pair (left) and a scene-centric image-caption pair (right). Note how the pairs differ considerably in terms of the visual style and the content of the caption. The pair on the left focuses on a single object (“pants”) and describes its fine-grained visual attributes (“multicolor,” “boho,” “batic”). The pair on the right captures a scene describing multiple objects (“seagulls,” “pier,” “people”) and relations between them (“sitting,” “watching”).

Research goals. We focus on (traditional) CMR methods that extract features from each modality and learn a common representation space. Recent years have seen extensive experimentation with such CMR methods, mostly organized into two groups: (i) contrastive experiments on object-centric datasets (Han et al., 2017), and (ii) contrastive experiments on scene-centric datasets (Lin et al., 2014). In this chapter, we consider representative state-of-the-art CMR methods from both groups. We select two pre-trained models which demonstrate state-of-the-art performance on CMR task and evaluate them in a zero-shot setting. In line with designs used in prior reproducibility work on CMR we select two models for the study. Following the ACM terminology (ACM, 2020), we focus on *reproducibility* (different team, same experimental setup) and *replicability* (different team, different experimental setup) of previously reported results. And following Voorhees (2002), we focus on relative (a.k.a. comparative) performance results. In addition, for the reproducibility experiment, we consider the absolute difference between the reported scores and the reproduced scores.

We address the following research questions: **(RQ1.1)** Are published relative performance results on CMR reproducible? This question matters because it allows us to confirm the validity of reported results. We show that the relative performance results are not fully reproducible. Specifically, the results are reproducible for one dataset, but not for the other dataset.

We then shift to replicability and examine whether lessons learned on scene-centric datasets transfer to object-centric datasets: **(RQ1.2)** To what extent are the published relative performance results replicable? That is, we investigate the validity of the reported results when evaluated in a different setup. We find that relative performance results are partially replicable, using other datasets.

After investigating the reproducibility and replicability of the results, we consider the generalizability of the results. We contrastively evaluate the results on object-centric and scene-centric datasets: **(RQ1.3)** Do relative performance results for state-of-the-art CMR methods generalize from scene-centric datasets to object-centric datasets? We discover that the relative performance results only partially generalize across the two dataset types.

Main contributions. Our main contributions are: (i) We are one of the first to consider reproducibility in the context of CMR and reproduce scene-centric CMR experiments from two papers (Radford et al., 2021; Zeng et al., 2022) and find that the results are only partially reproducible. (ii) We perform a replicability study and examine whether relative performance differences reported for CMR methods generalize from scene-centric datasets to object-centric datasets. (iii) We investigate the generalizability of obtained results and analyze the effectiveness of pre-training on scene-centric datasets for improving the performance of CMR on object-centric datasets, and vice versa. And, finally, (iv) to facilitate the reproducibility of our work, we provide the code and the pre-trained models used in our experiments

2.2 RELATED WORK

2.2.1 Cross-Modal Retrieval

CMR methods attempt to construct a multimodal representation space, where the similarity of concepts from different modalities can be measured. Some of the earliest approaches in CMR utilised canonical correlation analysis (Gong et al., 2014; Klein et al., 2014). They were followed by a dual encoder architecture equipped with a recurrent and a convolutional component, a hinge loss (Frome et al., 2013; Wang et al., 2016b) and hard-negative mining (Faghri et al., 2018). Later on, several attention-based architectures were introduced such as architectures with dual attention (Nam et al., 2017), stacked cross-attention (Lee et al., 2018), bidirectional focal attention (Liu et al., 2019).

Another line of work proposed to use transformer encoders (Vaswani et al., 2017) for CMR task (Messina et al., 2021), and adapted the BERT model (Devlin et al., 2019) as a backbone (Gao et al., 2020; Zhuge et al., 2021). Some other researchers worked on

improving CMR via modality-specific graphs (Wang et al., 2021b), or image and text generation modules (Gu et al., 2018).

There is also more domain-specific work that focused on CMR in fashion (Laenen et al., 2018; Laenen et al., 2017; Goei et al., 2021; Laenen, 2022), e-commerce (Hendriksen, 2022), cultural heritage (Sheng et al., 2021b) and cooking (Wang et al., 2021b).

In contrast to the majority of prior work on the topic, we focus on the reproducibility, replicability, and generalizability of CMR methods. In particular, we explore the state-of-the-art models designed for the CMR task by examining their performance on scene-centric and object-centric datasets.

2.2.2 Scene-Centric and Object-Centric Datasets

The majority of prior work related to object-centric and scene-centric datasets focuses on computer vision tasks such as object recognition, object classification, and scene recognition. Herranz et al. (2016) investigated biases in a CNN when trained on scene-centric versus object-centric datasets and evaluated on the task of object classification.

In the context of object detection, prior work focused on combining feature representations learned from object-centric and scene-centric datasets to improve the performance when detecting small objects (Shen et al., 2019), and using object-centric images to improve the detection of objects that do not appear frequently in complex scenes (Zhang et al., 2021a). Finally, for the task of scene recognition, Zhou et al. (2014) explored the quality of feature representations learned from both scene-centric and object-centric datasets and applied them to the task of scene recognition.

Unlike prior work on the topic, in this chapter, we focus on both scene-centric and object-centric datasets for evaluation on CMR task. In particular, we explore how SOTA CMR models perform on object-centric and scene-centric datasets.

2.2.3 Reproducibility in Cross-Modal Retrieval

To the best of our knowledge, despite the popularity of the CMR task, there are very few papers that focus on reproducibility of research in CMR. Some rare (recent) examples include Rao et al. (2022), where the authors analyze contributing factors that affect the performance of the state-of-the-art CMR models. However, all prior work focuses on exploring model performance only on two popular scene-centric datasets: Microsoft COCO (MS COCO) (Lin et al., 2014) and Flickr30k (Young et al., 2014).

In contrast, in this chapter, we take advantage of the diversity of the CMR datasets and specifically focus on examining how the state-of-the-art CMR models perform across different dataset types: scene-centric and object-centric datasets.

2.3 TASK DEFINITION

We follow the same notation as in previous work (Zhang et al., 2022c; Varamesh et al., 2020; Brown et al., 2020). An image-caption cross-modal dataset consists of a set of images \mathcal{I} and texts \mathcal{T} where the images and texts are aligned as image-text pairs: $\mathcal{D} = \{(\mathbf{x}_{\mathcal{I}}^1, \mathbf{x}_{\mathcal{T}}^1), \dots, (\mathbf{x}_{\mathcal{I}}^n, \mathbf{x}_{\mathcal{T}}^n)\}$.

The *cross-modal retrieval* (CMR) task is defined analogous to the standard information retrieval task: given a query \mathbf{q} and a set of m candidates $\Omega_{\mathbf{q}} = \{\mathbf{x}^1, \dots, \mathbf{x}^m\}$ we aim to rank all the candidates w.r.t. their relevance to the query \mathbf{q} . In CMR, the query can be either a text $\mathbf{q}_{\mathcal{T}}$ or an image $\mathbf{q}_{\mathcal{I}}$: $\mathbf{q} \in \{\mathbf{q}_{\mathcal{T}}, \mathbf{q}_{\mathcal{I}}\}$. Similarly, the set of candidate items can be either visual $\mathcal{I}_{\mathbf{q}} \subset \mathcal{I}$, or textual $\mathcal{T}_{\mathbf{q}} \subset \mathcal{T}$ data: $\Omega \in \{\mathcal{I}_{\mathbf{q}}, \mathcal{T}_{\mathbf{q}}\}$.

The CMR task is performed across modalities, therefore, if the query is a text then the set of candidates are images, and vice versa. Hence, the task comprises effectively two subtasks: (i) *text-to-image retrieval*: given a textual query $\mathbf{q}_{\mathcal{T}}$ and a set of candidate images $\Omega \subset \mathcal{I}$, we aim to rank all instances in the set of candidate items Ω w.r.t. their relevance to the query $\mathbf{q}_{\mathcal{T}}$; (ii) *image-to-text retrieval*: given an image as a query $\mathbf{q}_{\mathcal{I}}$ and a set of candidate texts $\Omega \subset \mathcal{T}$, we aim to rank all instances in the set of candidate items Ω w.r.t. their relevance to the query $\mathbf{q}_{\mathcal{I}}$.

2.4 METHODS

In this section, we give an overview of the models included in the study, of the models which were excluded, and provide justification for it. All the approaches we focus on belong to the traditional CMR framework and comprise two stages. First, we extract textual and visual features. The features are typically extracted with a textual encoder and a visual encoder. Next, we learn a latent representation space where the similarity of items from different modalities can be measured directly.

2.4.1 Methods Included for Comparison

We focus on CMR in *zero-shot setting*, hence, we only consider pre-trained models. Therefore, we focus on the models that are released for public use. Besides, as explained in Section 6.1, we follow prior reproducibility work to inform our experimental choices regarding the number of models. Given the above-mentioned requirements, we selected two methods that demonstrate state-of-the-art performance on the CMR task: CLIP and X-VLM.

Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021). This model

is a dual encoder that comprises an image encoder, and a text encoder. The model was pre-trained in a contrastive manner using a symmetric loss function. It is trained on 400 million image-caption pairs scraped from the internet. The text encoder is a transformer (Vaswani et al., 2017) with modification from (Radford et al., 2019). For the image encoder, the authors present two architectures. The first one is based on ResNet (He et al., 2016) and it is represented in five variants in total. The first two options are ResNet-50, ResNet-101; the last three options are variants of ResNet scaled up in the style of EfficientNet (Tan and Le, 2019). The second image encoder architecture is a vision transformer (ViT) (Dosovitskiy et al., 2021). It is presented in three variants: ViT-B/32, a ViT-B/16, and a ViT-L/14. The CMR results reported in the original paper are obtained with a model configuration where vision transformer ViT-L/14 is used as an image encoder, and the text transformer is a text encoder. Hence, we use this configuration in our experiments.

X-VLM (Zeng et al., 2022). This model consists of three encoders: an image encoder, a text encoder, and a cross-modal encoder. The image and text encoder take an image and text as inputs and output their visual and textual representations. The cross-modal encoder fuses the output of the image encoder and the output of the text encoder. The fusion is done via a cross-attention mechanism. For CMR task, the model is fine-tuned via a contrastive learning loss and a matching loss. All encoders are transformer-based. The image encoder is a ViT initialised with Swin Transformer_{base} (Liu et al., 2021). Both the text encoder and the cross-modal encoder are initialised using different layers of BERT (Devlin et al., 2019): the text encoder is initialized using the first six layers, whereas the cross-modal encoder is initialised using the last six layers.

2.4.2 Methods Excluded from Comparison

While selecting the models for the experiments, we considered other architectures with promising performance on the MS COCO and the Flickr30k datasets. Below, we outline the architectures we considered and explain why they were not included.

Several models such as Visual N-Grams (Li et al., 2017), Unicoder-VL (Li et al., 2020a), ViLT-B/32 (Kim et al., 2021), UNITER (Chen et al., 2020b) were excluded because they were consistently outperformed by CLIP on the MS COCO and Flickr30k datasets by large margins. Besides, we excluded ImageBERT (Qi et al., 2020) because it was outperformed by CLIP on the MS COCO dataset. ALIGN (Jia et al., 2021), AL-BEF (Li et al., 2021a), VinVL (Zhang et al., 2021b), METER (Dou et al., 2022) were not included because X-VLM consistently outperformed them. UNITER (Chen et al., 2020b) was beaten by both CLIP and X-VLM. We did not include other well-performing models such as ALIGN (Jia et al., 2021), Flamingo (Alayrac et al., 2022), CoCa (Yu et al.,

2022) because the pre-trained models were not publicly available.

2.5 EXPERIMENTAL SETUP

In this section, we discuss our experimental design including the choice of datasets, subtasks, metrics, and implementation details.

2.5.1 Datasets

We run experiments on two scene-centric and three object-centric datasets. Below, we discuss each of the datasets in more detail.

Scene-centric datasets. We experiment with two scene-centric datasets: (i) Microsoft COCO (MS COCO) (Lin et al., 2014) contains 123,287 images depicting regular scenes from everyday life with multiple objects placed in their natural contexts. There are 91 different object types such as “person”, “bicycle”, “apple”. (ii) Flickr30k (Young et al., 2014) contains 31,783 images of regular scenes from everyday life, activities, and events. For both scene-centric datasets, we use the splits provided in (Karpathy and Li, 2015). The MS COCO dataset is split into 113,287 images for training, 5,000 for testing and 5,000 for validation; the Flickr30k dataset has 29,783 images for training, 1,000 for testing and 1,000 for validation. In both datasets, every image was annotated with five captions using Amazon Mechanical Turk. Besides, we select one caption per image randomly and use the test set for our experiments.

Object-centric datasets. We consider three object-centric datasets in our experiments: (i) Caltech-UCSD Birds 200 (CUB-200) (Welinder et al., 2010) contains 11,788 images of 200 birds species. Each image is annotated with a fine-grained caption from (Reed et al., 2016). We selected one caption per image randomly. Each caption is at least 10 words long and does not contain any information about the birds’ species or actions. (ii) Fashion200k (Han et al., 2017) contains 209,544 images that depict various fashion items in five product categories (dress, top, pant, skirt, jacket) and their corresponding descriptions. (iii) Amazon Berkley Objects (ABO) (Collins et al., 2022) contains 147,702 product listings associated with 398,212 images. This dataset was derived from Amazon.com product listings. We selected one image per listing and used the associated product description as its caption. The majority of images depict a single product on a white background. The product is located in the center of the image and takes at least 85% of the image area. For all object-centric datasets, we use the splits provided by the dataset authors and use the test split for our experiments.

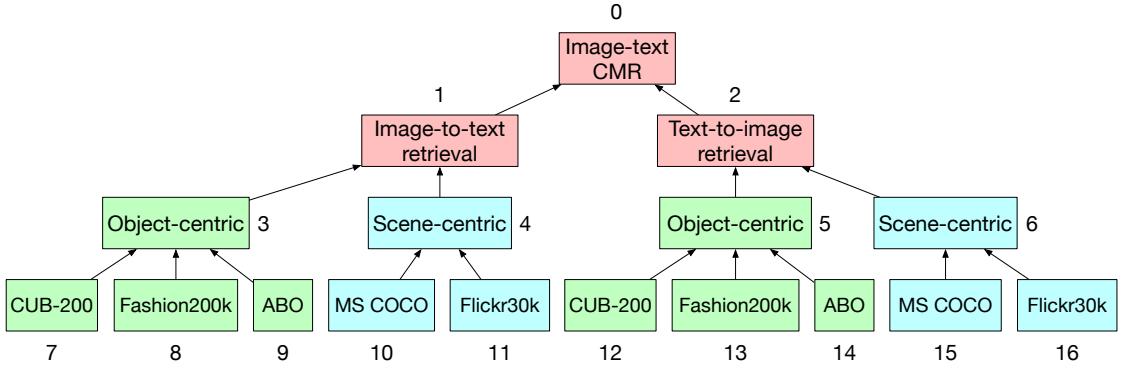


Figure 2.2: Our experimental design for evaluating CMR methods across object-centric and scene-centric datasets. The blue colour indicates parts of the tree used in Experiment 1, the green color indicates parts of the tree used in Experiment 2, and the red color indicates parts used in all experiments. (Best viewed in color.)

2.5.2 Subtasks

Our goal is to assess and compare the performance of the CMR methods (described in Section 2.4) across the object-centric and scene-centric datasets described in the previous subsection. We design an experimental setup that takes into account two CMR subtasks and two dataset types. It can be summarized using a tree with branches that correspond to different configurations (see Figure 2.2). We explain how we cover the branches of this tree in the next subsection.

The tree starts with a root (“Image-text CMR” with label 0) that has sixteen descendants, in total. The root node has two children corresponding to the two image-text CMR subtasks: text-to-image retrieval (node 1) and image-to-text retrieval (node 2). Since we want to evaluate each of these subtasks on both object-centric and scene-centric datasets, nodes 1 and 2 also have two children each, i.e., the nodes {3,4,5,6}. Finally, every object-centric node has three children: CUB-200, Fashion200k, and ABO datasets {7,8,9,12,13,14}; and every scene-centric node has two children: MS COCO and Flickr30k datasets {10,11,15,16}.

2.5.3 Experiments

To answer the research questions introduced in Section 6.1, we conduct two experiments. In all the experiments, we use CLIP and X-VLM models in a zero-shot setting. Following (Voorhees, 2002), we focus on relative performance results. In each experiment, we consider different subtrees from Figure 2.2. Following (Radford et al., 2021; Zeng et al., 2022; Li et al., 2017; Kim et al., 2021), we use Recall@K where $K = \{1, 5, 10\}$ to evaluate the model performance in all our experiments. In addition, following (Ueki,

2021; Zhang et al., 2022b; Song and Choi, 2021), we calculate the sum of recalls (rsum) for text-to-image, and image-to-text retrieval tasks as well as the total sum of recalls for both tasks.

For text-to-image retrieval, we first obtain representations for all the candidate images by passing them through the image encoder of the model. Then we pass each textual query through the text encoder of the model and retrieve the top- k candidates ranked by cosine similarity w.r.t. the query.

For image-to-text retrieval, we do the reverse, using the texts as candidates and images as queries. More specifically, we start by obtaining representations of the candidate captions by passing them through the text encoder. Afterwards, for each of the visual queries, we pass the query through the image encoder and retrieve top- k candidates ranked by cosine similarity w.r.t. the query.

In *Experiment 1* we evaluate the reproducibility of the CMR results reported in the original publications (RQ1.1). Both models we consider (CLIP and X-VLM) were originally evaluated on two scene-centric datasets, viz. MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). Therefore, for our reproducibility study, we also evaluate these models on these two datasets. We evaluate both text-to-image and image-to-text retrieval. That is, we focus on the two sub-trees $o \leftarrow 1 \leftarrow 4 \leftarrow \{10, 11\}$ and $o \leftarrow 2 \leftarrow 6 \leftarrow \{15, 16\}$ (the red and blue parts of the tree) from Figure 2.2. In addition to relative performance results, we consider absolute differences between the reported scores and the reproduced scores. Following Petrov and Macdonald (2022), we assume that the score is reproduced if we obtain a score value equal to the reported score given a relative tolerance of $\pm 5\%$.

In *Experiment 2* we focus on the replicability of the reported results on object-centric datasets (RQ1.2). Thus, we evaluate CLIP and X-VLM on the CUB-200 (Welinder et al., 2010), Fashion200k (Han et al., 2017), and ABO (Collins et al., 2022) datasets. This experiment covers the subtrees $o \leftarrow 1 \leftarrow 3 \leftarrow \{7, 8, 9\}$ and $o \leftarrow 2 \leftarrow 5 \leftarrow \{12, 13, 14\}$ (the red and green parts of the tree) in Figure 2.2.

After obtaining the results from Experiment 1 and 2, we examine the generalizability of the obtained scores (RQ1.3). We do so by comparing the relative performance results the models achieve on the object-centric versus scene-centric datasets. More specifically, we compare the relative performance of CLIP and X-VLM on CUB-200 (Welinder et al., 2010), Fashion200k (Han et al., 2017), ABO (Collins et al., 2022) with their relative performance on MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). Thus, this experiment captures the complete tree in Figure 2.2.

Table 2.1: Results of Experiment 1 (reproducibility study), using the MS COCO and Flickr30k datasets . “Orig.” indicates the scores from the original publications. “Repr.” indicates the scores that we obtained.

Model	Text-to-image			Image-to-text			Rsum			
	R@1	R@5	R@10	R@1	R@5	R@10	t2i	i2t	total	
MS COCO (5k)										
Orig.	CLIP (Radford et al., 2021)	37.80	62.40	72.20	58.40	81.50	88.10	172.40	228.00	400.40
	X-VLM (Zeng et al., 2022)	55.60	82.70	90.00	70.80	92.10	96.50	228.30	259.40	487.70
Repr.	CLIP	21.59	40.22	49.80	24.36	44.13	53.41	111.61	121.90	233.51
	X-VLM	42.79	67.61	67.64	64.60	84.48	84.50	178.04	233.58	411.62
Flickr30k (1k)										
Orig.	CLIP	68.70	90.60	95.20	88.00	98.70	99.40	254.50	286.10	540.60
	X-VLM	71.90	93.30	96.40	85.30	97.80	99.60	261.60	282.70	544.30
Repr.	CLIP	74.95	93.09	96.15	77.02	94.18	96.84	264.19	268.04	532.23
	X-VLM	37.82	82.36	82.48	63.30	91.10	91.10	202.66	245.50	448.16

2.6 RESULTS

We focus on the reproducibility (different team, same setup) and replicability (different team, different setup) of the CMR experiments reported in the original papers devoted to CLIP (Radford et al., 2021) and X-VLM (Zeng et al., 2022). To organize our result presentation, we refer to the tree in Figure 2.2. We traverse the tree bottom up, from the leaves to the root.

2.6.1 RQ1.1: Reproducibility

To address RQ1.1, we report on the outcomes of Experiment 1. We investigate to what extent the CMR results reported in the original papers devoted to CLIP (Radford et al., 2021) and X-VLM (Zeng et al., 2022) are reproducible. Given that both methods were originally evaluated on two scene-centric datasets, viz. MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014), we evaluate the models on the text-to-image and image-to-text tasks on these two datasets. Therefore, we focus on the two blue sub-trees $o \leftarrow 1 \leftarrow 4 \leftarrow \{10, 11\}$ and $o \leftarrow 2 \leftarrow 6 \leftarrow \{15, 16\}$ from Figure 2.2.

Results. The results of Experiment 1 are shown in Table 2.1. We recall the scores obtained in the original papers (Radford et al., 2021; Zeng et al., 2022) (“Orig.”) and the scores that we obtained (“Repr.”), on the MS COCO and Flickr30k datasets. Across the board, the scores that we obtained (the “reproduced scores”) tend to be lower than the scores obtained in the original publications (the “original scores”).

On the MS COCO dataset, X-VLM consistently outperforms CLIP, both in the original publications and in our setup, for both the text-to-image and the image-to-text

tasks. Moreover, this holds for all $R@n$ metrics, and, hence, for the Rsum metrics. Interestingly, the relative gains that we obtain tend to be larger than the ones obtained in the original publications. For example, our biggest relative difference is for the image-to-text task in terms of the $R@1$ metric: according to the scores reported in (Zeng et al., 2022; Radford et al., 2021), X-VLM outperforms CLIP by 21%, whereas in our experiments the relative gain is 165%.

On average, the original CLIP scores are as much as $\sim 70\%$ higher than the reproduced scores; the original scores for X-VLM are $\sim 20\%$ higher than the reproduced ones. When considering the absolute differences between the original scores and the reproduced scores and assuming a relative tolerance of $\pm 5\%$, we see that, on the MS COCO dataset, the scores are not reproducible for both models.

On the Flickr30k dataset, we see a different pattern. For the text-to-image task, the original results indicate that X-VLM consistently outperforms CLIP, on all $R@n$ metrics, but according to our results, the relative order is consistently reversed. For the image-to-text task, we obtained mixed outcomes: for $R@1$ and $R@5$, the original order (CLIP outperforms X-VLM) is confirmed, but for $R@10$ the order is swapped. According to our experimental results, however, CLIP consistently outperforms X-VLM on all tasks, and on all $R@n$ metrics (and hence also on the Rsum metrics).

On the Flickr30k dataset, the CLIP scores are reproduced on the text-to-image and image-to-text retrieval tasks when the model is evaluated on $R@5$ and $R@10$. On the text-to-image task, the reproduced $R@5$ score is 2.7% higher than the original score; the reproduced $R@10$ score is 1% higher than the original score. For the image-to-text retrieval task, the reproduced $R@5$ score is 4% lower than the original score; the reproduced $R@10$ score is 2% lower than the original score.

Answer to RQ1.1. In the case of the CLIP model, the obtained *absolute* scores were reproducible only on the Flickr30k dataset for the text-to-image and the image-to-text tasks when evaluated on $R@5$ and $R@10$. For X-VLM, we did not find the absolute scores obtained when evaluating the model on the MS COCO and Flickr20k datasets to be reproducible, neither for the text-to-image nor the image-to-text tasks.

The *relative* outcomes on the MS COCO dataset could be reproduced, for all tasks and metrics, whereas on the Flickr30k dataset, they could only partially be reproduced, that is, only for the image-to-text task on the $R@1$ and $R@5$ metrics; for the text-to-image task, X-VLM outperforms CLIP according to the original scores, but CLIP outperforms X-VLM according to our reproduced scores.

Upshot. As explained in Section 2.4, in this chapter we focus on CMR in a zero-shot setting. This implies that the differences that we observed between the original scores and the reproduced scores must be due to differences in text and image data processing and loading. We, therefore, recommend that the future work includes (as much as is practically possible) tools and scripts used in these stages of the experiment

with the publication of its implementations.

2.6.2 RQ1.2: Replicability

To answer RQ1.2, we replicate the originally reported text-to-image and image-to-text retrieval experiments in a different setup, i.e., by evaluating CLIP and X-VLM using object-centric datasets instead of scene-centric datasets. Thus, we evaluate CLIP and X-VLM on the CUB-200 (Welinder et al., 2010), Fashion200k (Han et al., 2017), and ABO (Collins et al., 2022) datasets and focus on the green subtrees $o \leftarrow 1 \leftarrow 3 \leftarrow \{7, 8, 9\}$ and $o \leftarrow 2 \leftarrow 5 \leftarrow \{12, 13, 14\}$ from Figure 2.2.

Results. The results of Experiment 2 (aimed at answering RQ1.2) can be found in Table 2.2. On the CUB-200 (Welinder et al., 2010) dataset, CLIP consistently outperforms X-VLM. The biggest relative increase is 124% for image-to-text in terms of R@10, while the smallest relative increase is 1% for text-to-image in terms of R@1. Overall, on the text-to-image retrieval task, CLIP outperforms X-VLM by 38%, and on the image-to-text retrieval task, the relative gain is 70%.

On Fashion200k (Han et al., 2017), CLIP outperforms X-VLM, too. The smallest relative increase is 9% for text-to-image in terms of R@1, and the biggest relative increase is 260% for image-to-text in terms of R@10. In general, on the text-to-image retrieval task, CLIP outperforms X-VLM by 52%; on the image-to-text retrieval task, the relative gain is 83%.

Finally, on the ABO (Collins et al., 2022) dataset, CLIP outperforms X-VLM again. The smallest relative increase is 101% for text-to-image in terms of R@1, and the biggest relative increase is 241% for image-to-text again in terms of R@10. In general, on the text-to-image retrieval task, CLIP outperforms X-VLM by 139%; on the image-to-text retrieval task, the relative gain is 190%. All in all, CLIP outperforms X-VLM on all three scene-centric datasets. The overall relative gain on CUB-200 (Welinder et al., 2010) dataset is 55%, on Fashion200k (Han et al., 2017) dataset – 101%. The biggest relative gain of 166% is obtained on the ABO (Collins et al., 2022) dataset.

Answer to RQ1.2. The outcome of Experiment 2 is clear. The original relative performance results obtained on the MS COCO and Flickr30k (Table 2.1) are only partially replicable to the CUB-200, Fashion200k, and ABO datasets. On the latter datasets CLIP consistently outperforms X-VLM by a large margin, whereas the original scores obtained on the former datasets indicate that X-VLM mostly outperforms CLIP.

Upshot. We hypothesize that the failure to replicate the relative results originally reported for scene-centric datasets (viz. X-VLM outperforms CLIP) is due to CLIP being pre-trained on more and more diverse image data. We, therefore, recommend that future work aimed at developing large-scale CMR models quantifies and reports

Table 2.2: Results of Experiment 2 (replicability study), using the CUB-200, Fashion200k, and ABO datasets.

Model	Text-to-image			Image-to-text			Rsum		
	R@1	R@5	R@10	R@1	R@5	R@10	t2i	i2t	total
CUB-200									
CLIP	0.71	2.38	4.42	1.23	3.40	5.48	7.51	10.11	17.62
X-VLM	0.70	2.28	2.45	1.16	2.35	2.45	5.43	5.96	11.39
Fashion200k									
CLIP	3.05	8.56	12.85	3.43	9.82	14.56	24.46	27.81	52.27
X-VLM	2.80	6.62	6.70	1.84	3.96	4.04	16.12	09.84	25.96
ABO									
CLIP	6.25	13.90	18.50	7.99	18.96	25.57	38.65	52.52	91.17
X-VLM	3.10	6.48	6.56	3.20	7.42	7.50	16.14	18.12	34.26

the diversity of the training data used.

2.6.3 RQ1.3: Generalizability

To answer RQ1.3, we compare the relative performance of the selected models on object-centric and scene-centric data. Thus, we compare the relative performance of CLIP and X-VLM on CUB-200 (Welinder et al., 2010), Fashion200k (Han et al., 2017), ABO (Collins et al., 2022) with their relative performance on MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). We focus on the complete tree from Figure 2.2.

Results. The results of our experiments on the scene-centric datasets are in Table 2.1; the results that we obtained on the object-centric datasets are in Table 2.2. On object-centric datasets, CLIP consistently outperforms X-VLM. However, the situation with scene-centric results is partially the opposite. There, X-VLM outperforms CLIP on the MS COCO dataset.

Answer to RQ1.3. Hence, we answer RQ1.3 by stating that the relative performance results for CLIP and X-VLM that we obtained in our experiments only partially generalize from scene-centric to object-centric datasets. The MS COCO dataset is the odd one out.¹

Upshot. Given the observed differences in relative performance results for CLIP and X-VLM on scene-centric vs. object-centric datasets, we recommend that CMR be trained

¹ On the GitHub repository for CLIP, several issues have been posted related to the performance of CLIP on the MS COCO dataset. See, e.g., <https://github.com/openai/CLIP/issues/115>.

in both scene-centric and object-centric datasets to help improve the generalizability of experimental outcomes.

2.7 DISCUSSION AND CONCLUSION

We have examined two SOTA image-text CMR methods, CLIP and X-VLM, by contrasting their performance on two scene-centric datasets (MS COCO and Flickr30k) and three object-centric datasets (CUB-200, Fashion200k, ABO) in a zero-shot setting.

We focused on the *reproducibility* of the CMR results reported in the original publications when evaluated on the selected scene-centric datasets. The reported scores were not reproducible for X-VLM when evaluated on the MS COCO and the Flickr30k datasets. For CLIP, we were able to reproduce the scores on the Flickr30k dataset when evaluated using R@5 and R@10. Conversely, the relative results were reproducible on the MS COCO dataset, for all metrics and tasks, and partially reproducible on the Flickr30k dataset only for the image-to-text task when evaluated on R@1 and R@5. We also examined the *replicability* of the CMR results using three object-centric datasets. We discovered that the relative results are replicable when we compare the relative performance on the object-centric datasets with the relative scores on the Flickr30k dataset. However, for the MS COCO dataset, the relative outcomes were not replicable. And, finally, we explored the generalizability of the obtained results by comparing the models' performance on scene-centric vs. object-centric datasets. We observed that the absolute scores obtained when evaluating models on object-centric datasets are much lower than the scores obtained on scene-centric datasets.

Our findings demonstrate that the reproducibility of CMR methods on scene-centric datasets is an open problem. Besides, we show that while the majority of CMR methods are evaluated on the MS COCO and the Flickr30k datasets, the object-centric datasets represent a challenging and relatively unexplored set of benchmarks.

A limitation of our work is the relatively small number of scene-centric and object-centric datasets used for the evaluation of the models. Another limitation is that we only considered CMR in a zero-shot setting, ignoring, e.g., few-shot scenarios; this limitation did, however, come with the important advantage of reducing the number of experimental design decisions to be made for contrastive experiments.

A promising direction for future work is to include further datasets when contrasting the performance of CMR models, both scene-centric and object-centric. In particular, it would be interesting to investigate the models' performance on datasets, e.g., Conceptual Captions (Sharma et al., 2018), the Flower (Nilsback and Zisserman, 2008), and the Cars (Krause et al., 2013) datasets. A natural step after that would be to consider few-shot scenarios.

Thus, our answer to RQ1 is that while relative performance results in image-text CMR are partially reproducible and replicable across certain datasets, particularly scene-centric ones, they face challenges on object-centric datasets. The absolute performance scores on object-centric datasets are lower compared to scene-centric datasets, emphasising the need for further exploration and evaluation of CMR methods on diverse benchmark datasets. Reproducibility on scene-centric datasets is a challenge, with partial success attributed to differences in data preprocessing and experimental setups. Replicability from scene-centric to object-centric datasets is limited, indicating discrepancies in model performance due to dataset characteristics. Generalizability across different dataset types is constrained, with lower performance on object-centric datasets suggesting that current models lack robustness across diverse data types.

REPRODUCIBILITY

To ensure the reproducibility of the findings presented in this chapter, we have made our code publicly accessible at <https://github.com/mariyahendriksen/ecir23-object-centric-vs-scene-centric-CMR>.

3

MULTIMODAL LEARNED SPARSE RETRIEVAL

In the previous chapter, we focused on the problem of reproducibility of cross-modal retrieval (CMR) results for object-centric and scene-centric datasets. In this chapter, we continue our investigation of dense and sparse retrieval in the context of vision-language (VL) alignment in the context of learned sparse retrieval (LSR) (Formal et al., 2021; Zamani et al., 2018). LSR represents a promising research direction due to its potential for efficient and effective neural retrieval, particularly in text-based tasks. However, the application of LSR in VL context remains relatively underexplored (Nguyen et al., 2023b). Hence, in this chapter, we address this research question:

RQ2: How can learned sparse retrieval techniques be applied in the vision-language domain?

To answer this research question, we design and implement a model for the task and evaluate its performance. During our experiments, we discover two phenomena arising in the domain – dimension co-activation and semantic deviation. Dimension co-activation refers to the scenario where sparse representations of images and captions activate the same output dimensions, creating a subset of dense space within the vocabulary. While some co-activation is necessary for effectively matching captions with corresponding images, excessive co-activation can lead to inefficient retrieval processes. Semantic deviation, on the other hand, highlights the disparity between the semantic content of the visual or textual query and the sparse output terms. Upon discovering and formalizing the phenomena of dimension co-activation and semantic deviation, we formally define both and propose the metrics to quantify them. Furthermore, to mitigate both phenomena, we propose to train our model with probabilistic

This chapter was published at the 44th European Conference on Information Retrieval (ECIR 2024) under the title “Multimodal Learned Sparse Retrieval with Probabilistic Expansion Control” (Nguyen et al., 2024).

expansion control, gradually increasing term expansion toward the end of the training process. This chapter contributes to our understanding of the problem of VL alignment in the context of LSR.

3.1 INTRODUCTION

LSR (Formal et al., 2021; Formal et al., 2022; Nguyen et al., 2023b) typically employs transformer-based encoders to encode queries and documents into sparse lexical vectors (i.e., bags of weighted terms) that are compatible with a traditional inverted index. Empirically, LSR has shown advantages over single-vector dense models on retrieval generalization benchmarks (Formal et al., 2022; Kamalloo et al., 2023).

While LSR and dense retrieval are prevalent in text retrieval, dense retrieval has taken the lead in multi-modal search. This is evident in state-of-the-art text-image pretraining methods like BLIP (Li et al., 2022b) and ALBEF (Li et al., 2021a), which rely on dense architectures. For multimodal learned sparse retrieval (MLSR), LexLIP (Zhao et al., 2023a) and STAIR (Chen et al., 2023a) are the only recent methods that exhibit competitive results on standard benchmarks. However, both models require complex multi-step training on extensive text-image pairs: LexLIP with up to 14.3 million pairs and STAIR with a massive 1 billion pairs, encompassing public and private data.

We approach the MLSR problem by using a pre-trained dense model and training a small sparse projection head on top of dense vectors, using image-text dense scores as a supervision signal. Naively learning the projection layer leads to issues of (i) high dimension co-activation and (ii) semantic deviation. Issue (i) happens when text and image sparse vectors excessively activate the same output dimensions, forming a sub-dense space inside the vocabulary space. Issue (ii) means that produced output terms do not reflect the content of captions/images, making them not human-interpretable. To counter (i) and (ii), we propose a single-step training method with probabilistic term expansion control. By disabling term expansions, we force the projection to produce meaningful terms first, then gradually allow more term expansions to improve the effectiveness while also randomly reminding the model not to fully rely on expansion terms. This process is handled using Bernoulli random variables with a parameter scheduler to model the expansion likelihood at both caption and word levels.

Opting for dense to sparse projection, instead of training an MLSR model from scratch, provides several advantages. First, it is aligned with the broader community effort to reduce the carbon footprint of training deep learning models (Luccioni and Hernandez-Garcia, 2023). By keeping the dense encoders frozen and learning a light-weight projection layer, we can avoid the double GPU training/inference cost of two models (dense & sparse) while having more flexibility. Our approach enables the pre-computation of dense text and image vectors, allowing easy integration or removal of

the projection layer based on available (dense or sparse) software and infrastructure. Moreover, this transformation may shed light on the interpretability of dense vectors, possibly contributing to a deeper understanding of the fundamental distinctions between these two multi-modal retrieval paradigms.

To evaluate the proposed training method, we conduct extensive experiments on two dense multi-modal models (BLIP, ALBEF) and two scene-centric (Hendriksen et al., 2023) datasets (MSCOCO (Lin et al., 2014), Flickr30k (Young et al., 2014)). We analyze the problems of dimension co-activation and semantic deviation under different settings.

Our contributions. The main contributions of this chapter are: (i) We propose a line of research for efficiently converting a multi-modal dense retrieval model to a multi-modal LSR model. (ii) We train a lightweight projection head to convert dense to sparse vectors and show that our sparsified models are faithful to dense models while outperforming previous multi-modal LSR models. The training is efficient and does not require ground-truth labels. (iii) We identify the issues of high dimension co-activation and semantic deviation and propose a training method to address them.

3.2 RELATED WORK

3.2.1 Learned Sparse Retrieval

Learned sparse retrieval is a family of neural retrieval methods that encode queries and documents into sparse lexical vectors that can be indexed and searched efficiently with an inverted index. There are many LSR approaches in the literature on text retrieval (Formal et al., 2021; Zamani et al., 2018; Nguyen et al., 2023c); they are mainly built up from two types of encoder: MLP and MLM (Nguyen et al., 2023b). The MLP encoder uses a linear feedforward layer placed on top of the transformers’s last contextualized embeddings to predict the importance of input terms (similar to term-frequency in traditional lexical retrieval). The MLP encoder has no term expansion capability. On the other hand, the MLM encoder utilizes the logits of the masked language model (MLM) for weighting terms and selecting expansion terms. Splade (Formal et al., 2022; Formal et al., 2021) is a recent state-of-the-art text-oriented LSR approach that employs the MLM encoder in both query and document side, while other methods (MacAvaney et al., 2020; Lin and Ma, 2021; Dai and Callan, 2019) use MLP encoders on both sides or only on the query side. Although it seems to be more beneficial to have expansion on both queries and documents, a recent study (Nguyen et al., 2023b) found that query and document expansion have a cancellation effect on text retrieval (i.e., having expansion on the document side reduces the usefulness

of query expansion) and one could obtain near state-of-the-art results without query expansion.

Unlike prior work focused on converting sparse to dense representations for hybrid ad-hoc text retrieval (Lin and Lin, 2021; Lin and Lin, 2023), our work explores the reverse task of dense to sparse conversion in the multi-modal domain. This direction presents new challenges due to dimension co-activation and semantic deviation issues. Ram et al. (Ram et al., 2023) interpreted text dense retrieval by zero-shot projection from dense to vocabulary space using a frozen MLM head. Nguyen et al. (2023a) propose a simple sparse VL bi-encoder without query expansion and evaluate the performance on the image suggestion task. We aim for an efficient, effective, and semantically faithful drop-in sparse replacement of multi-modal dense retrieval, necessitating training of the projection layer.

3.2.2 Cross-Modal Retrieval

CMR methods construct a multimodal representation space, where the similarity of concepts from different modalities can be measured using a distance metric such as a cosine or Euclidean distance. Some of the earliest approaches in CMR utilized canonical correlation analysis (Gong et al., 2014; Klein et al., 2014). They were followed by a dual encoder architecture equipped with a recurrent and a convolutional component, the most prominent approaches in that area featured a hinge loss (Frome et al., 2013; Wang et al., 2016b). Later approaches further improved the effectiveness using hard-negative mining (Faghri et al., 2018).

Later, the integration of attention mechanisms improved performance. This family of attention mechanisms includes dual attention (Nam et al., 2017), stacked cross-attention (Lee et al., 2018), bidirectional focal attention (Liu et al., 2019). Another line of work proposes to use transformer encoders (Vaswani et al., 2017) for this task (Messina et al., 2021), and adapts the BERT model (Devlin et al., 2019) as a backbone (Gao et al., 2020; Zhuge et al., 2021).

A related line of work focuses on improving the performance on CMR via modality-specific graphs (Wang et al., 2021b), or image and text generation modules (Gu et al., 2018). There is also more domain-specific work that focuses on CMR in fashion (Goei et al., 2021; Laenen, 2022), e-commerce (Hendriksen et al., 2022; Hendriksen, 2022), cultural heritage (Sheng et al., 2021b), and cooking (Wang et al., 2021b).

3.3 BACKGROUND

Task definition. We use the same notation as in previous work (Zhang et al., 2022c; Brown et al., 2020). We work with a cross-modal dataset \mathcal{D} that includes N image-caption tuples: $\mathcal{D} = \left\{ \left(\mathbf{x}_{\mathcal{I}}^i, \{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k \right) \right\}_{i=1}^N$. Each tuple comprises an image $\mathbf{x}_{\mathcal{I}}$ and k associated captions $\{\mathbf{x}_{\mathcal{C}_j}\}_{j=1}^k$.

The *cross-modal retrieval* (CMR) task is defined analogously to the standard information retrieval task: given a query and a set of candidates, we rank all candidates w.r.t. their relevance to the query. The query can be either a caption or an image. Similarly, the set of candidate items can contain either images or captions. CMR is performed across modalities, therefore, if the query is a caption then the set of candidates are images, and vice versa. Hence, the task comprises two subtasks: (i) *caption-to-image retrieval*: retrieving images relevant to a caption query, and (ii) *image-to-caption retrieval*: retrieving relevant captions that describe an image query. We focus on *caption-to-image retrieval* only as it is more challenging, as reported by previous research (Li et al., 2022b; Li et al., 2021a; Zhao et al., 2023a).

Sparsification-induced phenomena. In this chapter, we investigate two phenomena arising during the sparsification process: dimension co-activation and semantic deviation.

Definition 1 (Dimension co-activation). We define *dimension co-activation* as sparse image and caption representations activating the same output dimensions, creating a sub-dense space within the vocabulary. While co-activation is essential for matching captions with images and can be measured by FLOPs, *high co-activation* results in unnecessarily long posting lists, harming the efficiency of LSR. Establishing a clear threshold for *high co-activation* is challenging, but we observe that beyond a certain point, increased FLOPs yield minimal improvements in effectiveness. To quantify this effect, we use effectiveness metrics (e.g., R@k) in combination with the FLOPs metric:

$$\text{FLOPs} = \frac{1}{|\mathcal{C}| |\mathcal{I}|} \sum_{\mathbf{x}_{\mathcal{C}} \in \mathcal{C}} \sum_{\mathbf{x}_{\mathcal{I}} \in \mathcal{I}} \mathbf{s}_{\mathcal{C}}^0 \cdot \mathbf{s}_{\mathcal{I}}^0, \quad (3.1)$$

where \mathcal{C} and \mathcal{I} are caption and image collections, $\mathbf{s}_{\mathcal{C}}, \mathbf{s}_{\mathcal{I}}$ are sparse vectors of a caption $\mathbf{x}_{\mathcal{C}}$ and an image $\mathbf{x}_{\mathcal{I}}$.

Definition 2 (Semantic deviation). We define *semantic deviation* as the disparity between the semantic information in the visual or textual query and that in the sparse output terms. High co-activation suggests (but does not guarantee) semantic deviation.

Measuring semantic deviation directly is challenging, so we use two rough proxies, *Exact@k* and *Semantic@k*, defined as follows:

$$\text{Exact}@k = \frac{1}{k} |\{t \mid t \in \mathbf{x}_{\mathcal{C}}, t \in \text{top}_k(\mathbf{s}_{\mathcal{C}})\}| \quad (3.2)$$

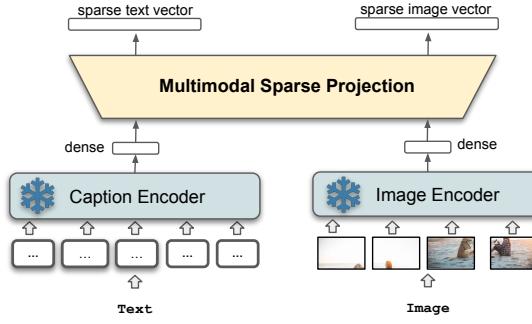


Figure 3.1: The architecture of Dense2Sparse (D2S). The caption and image encoders are frozen, and the sparse projection is trained to project dense vectors to sparse vectors.

$$\text{Semantic}@k = \frac{1}{k} \sum_{\mathbf{x}_t^i \in \text{top}_k(\mathbf{s}_C)} \max_{\mathbf{x}_t^j \in \mathbf{x}_C} \frac{f_{enc}(\mathbf{x}_t^i) \cdot f_{enc}(\mathbf{x}_t^j)}{\|f_{enc}(\mathbf{x}_t^i)\| \|f_{enc}(\mathbf{x}_t^j)\|}. \quad (3.3)$$

Exact@k measures the ratio of overlapping terms between the input caption and the top- k highest weighted output terms, providing a partial picture of semantic deviation without considering synonyms. *Semantic@k* complements *Exact@k* by calculating the averaged cosine similarity between static embeddings obtained using model $f_{enc}(\cdot)$ of top- k output terms and input caption terms. Higher values for both metrics suggest less semantic deviation, implying better alignment of output terms with input captions.

3.4 METHODOLOGY

3.4.1 Model Architecture

The architecture of our Dense2Sparse model is visualized in Figure 3.1.

Dense2Sparse takes an image and a caption as input, projecting them into a $|V|$ -dimensional space, where each dimension represents the weight of a corresponding vocabulary entry. The key components include two dense encoders, an image encoder $f_\theta^T(\cdot)$ and a caption encoder $f_\phi^C(\cdot)$, as well as a multimodal sparse projection head $g_\psi(\cdot)$.

Dense image and text encoders. The *dense image encoder* $f_\theta^T : \mathcal{X} \rightarrow \mathcal{Z}$ takes an input image \mathbf{x}_I and maps it into a latent space $\mathcal{Z} = \mathcal{R}^d$: $\mathbf{z}_I = f_\theta^T(\mathbf{x}_I)$, where $\mathbf{z}_I \in \mathcal{R}^d$. Similarly, the *dense text encoder* $f_\phi^C : \mathcal{X} \rightarrow \mathcal{Z}$ takes an input text (caption) \mathbf{x}_C , and maps it into a latent space $\mathcal{Z} = \mathcal{R}^d$: $\mathbf{z}_C = f_\phi^C(\mathbf{x}_C)$, where $\mathbf{z}_C \in \mathcal{R}^d$. We obtain dense representations using BLIP and ALBEF as a backbone. Both encoders are frozen.

Multimodal sparse projection head. The *multimodal sparse projection head* $g_\psi : \mathcal{Z} \rightarrow \mathcal{S}$

maps dense latent image and text representations into the sparse image and text vector space $\mathcal{S} = \mathcal{R}_{>0}^{|V|}$:

$$\mathbf{s}_C = g_\psi(\mathbf{z}_C) \quad \text{and} \quad \mathbf{s}_I = g_\psi(\mathbf{z}_I). \quad (3.4)$$

The multimodal sparse projection head comprises four steps. First, we project the d -dimensional dense vector \mathbf{z} to an ω -dimensional dense vector: $\mathbf{z}_1 = \mathbf{W}_1 \mathbf{z}$, where $\mathbf{W}_1 \in \mathcal{R}^{\omega \times d}$, $\mathbf{z} \in \mathcal{R}^d$, and $\mathbf{z}_1 \in \mathcal{R}^\omega$. Second, we apply layer normalization:

$$\mathbf{z}_2 = \frac{\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1]}{\sqrt{\text{Var}[\mathbf{z}_1] + \epsilon}} \cdot \gamma + \beta, \quad (3.5)$$

where $\mathbb{E}[\mathbf{z}_1]$ and $\text{Var}[\mathbf{z}_1]$ are the expectation and variance of \mathbf{z}_1 , γ and β are learnable affine transformation parameters, and $\mathbf{z}_2 \in \mathcal{R}^\omega$. Third, we project \mathbf{z}_2 to the vocabulary space $\mathcal{S} = \mathcal{R}_{>0}^{|V|}$: $\mathbf{s} = \mathbf{W}_2 \mathbf{z}_2$, where $\mathbf{W}_2 \in \mathcal{R}^{|V| \times \omega}$, $\mathbf{z}_2 \in \mathcal{R}^\omega$, and $\mathbf{s} \in \mathcal{R}^{|V|}$. \mathbf{W}_2 is initialized with vocabulary embeddings similar to the transformer-masked language model. Fourth, we remove negative weights and apply a logarithmic transformation to the positive weights: $\mathbf{s} = \log_e(1 + \max(0, \mathbf{s}))$, where $\mathbf{s} \in \mathcal{R}_{>0}^{|V|}$. The resulting $|V|$ -dimensional sparse vector is aligned with the vocabulary, and each dimension represents the weight of the corresponding vocabulary entry. This projection head is similar to the MLM head employed in previous work (Formal et al., 2022; MacAvaney et al., 2020).

Probabilistic expansion control. Without any intervention, training the projection module with a standard contrastive loss could lead to high-dimension co-activation and semantic deviation as defined previously. This phenomenon affects the efficiency of an inverted index and the interpretability of the outputs. To mitigate this problem, we propose a single-step training algorithm with probabilistic lexical expansion control. It is described in Algorithm 1.

We define a Bernoulli random variable $\mathcal{E} \sim \text{Ber}(p)$, $p \in [0, 1]$ and use it to control textual query expansion. We consider a caption-level and a word-level expansion. The *caption-level expansion* is controlled by the random variable $\mathcal{E}_C \sim \text{Ber}(p_C)$. If $\mathcal{E}_C = 1$ the expansion is allowed, while $\mathcal{E}_C = 0$ means the expansion is not allowed. Analogously, the *word-level expansion*, or the expansion to the i -th word in the vocabulary, is regulated by the random variable $\mathcal{E}_i^v \sim \text{Ber}(p_i^v)$.

The parameters p_C and p_i^v define the likelihood of caption-level and word-level expansion within a given training epoch. During training, we initially set the caption-level expansion probability, p_C , to zero. This initial value prevents the expansion of textual queries, forcing the model to project images onto relevant tokens belonging to the captions they were paired with. This approach facilitates the meaningful projection of dense vectors onto relevant words in the vocabulary. However, it adversely impacts retrieval effectiveness, as the model cannot expand queries. As a consequence,

Algorithm 1 Multimodal LSR training with probabilistic expansion control

Input: image-caption pair ($\mathbf{x}_{\mathcal{I}}$, $\mathbf{x}_{\mathcal{C}}$), caption encoder $f_{\phi}^{\mathcal{C}}$, image encoder $f_{\theta}^{\mathcal{I}}$, sparse projection head g_{ψ} , loss function \mathcal{L} , and expansion rate function f_{incr} .

```

 $p_i^v \leftarrow 1 - df_i^v$ 
 $p_c \leftarrow 0$ 

for epoch do
    for batch do
         $\mathbf{z}_{\mathcal{C}} \leftarrow f_{\phi}^{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$ ,  $\mathbf{z}_{\mathcal{I}} \leftarrow f_{\theta}^{\mathcal{I}}(\mathbf{x}_{\mathcal{I}})$ 
         $\mathbf{s}_{\mathcal{C}} \leftarrow g_{\psi}(\mathbf{z}_{\mathcal{C}})$ ,  $\mathbf{s}_{\mathcal{I}} \leftarrow g_{\psi}(\mathbf{z}_{\mathcal{I}})$ 
         $\mathcal{E}_{\mathcal{C}} \sim \text{Ber}(p_c)$ ,  $\mathcal{E}_i^v \sim \text{Ber}(p_i^v)$ 
         $\bar{\mathbf{s}}_{\mathcal{C}} \leftarrow \text{EXPAND}(\mathbf{x}_{\mathcal{C}}, \mathbf{s}_{\mathcal{C}}, \mathcal{E}_{\mathcal{C}}, \mathcal{E}_i^v)$ 
         $\mathcal{L} \leftarrow \mathcal{L}(\bar{\mathbf{s}}_{\mathcal{C}}, \mathbf{s}_{\mathcal{I}}, \mathbf{z}_{\mathcal{I}}, \mathbf{z}_{\mathcal{C}})$ 
    end for
     $p_c \leftarrow f_{\text{incr}}(p_c)$ ,  $p_i^v \leftarrow f_{\text{incr}}(p_i^v)$ 
end for

```

the model’s ability to handle semantic matching is limited. To gradually relax this constraint, we use a scheduler that incrementally increases the value of p after each epoch until it reaches a maximum value of one in the final epoch. In each epoch, we sample the values of \mathcal{E} per batch and enforce expansion terms to be zero when $\mathcal{E}_{\mathcal{C}}$ equals zero. Similarly, for word-level expansion, we initialize the expansion probability of the i -th word p_i^v to $1 - df_i^v$ where df_i^v is the normalized document frequency of vocabulary element v_i in the caption collection \mathcal{C} . This setting discourages the expansion of more frequent terms because they are less meaningful and can hinder the efficiency of query processing algorithms. We relax each p_i^v after every epoch, ensuring that it reaches a maximum value of one at the conclusion of the training process. The expansion rate increase after each epoch is defined as follows:

$$f_{\text{incr}}(p) = \begin{cases} p + \frac{1}{\# \text{epochs}}, & \text{for caption-level expansion} \\ p + \frac{df_i^v}{\# \text{epochs}}, & \text{for word-level expansion.} \end{cases} \quad (3.6)$$

3.4.2 Training Loss

We train our Dense2Sparse using a loss that represents a weighted sum of a bidirectional loss and a sparse regularization parameter. The bidirectional loss is based on the following one-directional loss:

$$\ell^{(\mathcal{A} \rightarrow \mathcal{B})} = - \left(\frac{\exp(\mathbf{z}_\mathcal{A}^\top \mathbf{z}_\mathcal{B} / \tau)}{\sum_{\mathcal{I}^*} \exp(\mathbf{z}_\mathcal{A}^\top \mathbf{z}_{\mathcal{I}^*} / \tau)} \right) \log_2 \left(\text{SoftMax}[\mathbf{s}_\mathcal{A}^\top \mathbf{s}_\mathcal{B}] \right),$$

where $\mathbf{s}_\mathcal{A} \in \mathcal{R}_{>0}^{|\mathcal{V}|}$ and $\mathbf{s}_\mathcal{B} \in \mathcal{R}_{>0}^{|\mathcal{V}|}$ are sparse vectors, $\mathbf{z}_\mathcal{A} \in \mathcal{R}^d$ and $\mathbf{z}_\mathcal{B} \in \mathcal{R}^d$ are dense vectors, and $\tau \in \mathcal{R}_{>0}$ is a temperature parameter.

The resulting loss is formalized to capture both bidirectional losses and sparse regularization. The overall loss \mathcal{L} is defined as:

$$\mathcal{L} = (1 - \lambda) \underbrace{[\ell^{(\mathcal{I} \rightarrow \mathcal{C})} + \ell^{(\mathcal{C} \rightarrow \mathcal{I})}]}_{\text{bidirectional loss}} + \lambda \underbrace{\eta [L_1(\mathbf{s}_\mathcal{I}) + L_1(\mathbf{s}_\mathcal{C})]}_{\text{sparse regularization parameter}}, \quad (3.7)$$

where $\ell^{(\mathcal{I} \rightarrow \mathcal{C})}$ is an image-to-caption loss, $\ell^{(\mathcal{C} \rightarrow \mathcal{I})}$ is a caption-to-image loss; $\lambda = [0, 1]$ is a scalar weight, $\eta = [0, 1]$ is a sparsity regularization parameter, and $L_1(\mathbf{x}) = \|\mathbf{x}\|_1$ is L_1 regularization. It is worth noting that the loss utilizes dense scores for supervision, a strategy found to be more effective than using ground truth labels.

3.5 EXPERIMENTS AND RESULTS

3.5.1 Experimental Setup

Datasets. We trained and evaluated our models on two widely used datasets for text-image retrieval: MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015). Each image in the two datasets is paired with five short captions (with some exceptions). We re-used the splits from (Karpathy and Li, 2015) for training, evaluating, and testing. The splits on MSCOCO have 113.2k pairs for training, and 5k pairs for each validation/test set. Flickr30 is smaller with 29.8k/1k/1k for train, validation, test splits respectively. The best model is selected based on the validation set and evaluated on the test set.

Evaluation metrics. To evaluate model performance and effectiveness, we report R@k where $k = \{1, 5\}$, and MRR@10 using the *ir_measures* (MacAvaney et al., 2022) library.

Implementation and training details. The caption and image dense vectors of BLIP (Li et al., 2022b) and ALBEF (Li et al., 2021a) models are pre-computed with checkpoints from the larvis library (Li et al., 2022a). We train our models to convert from dense vectors to sparse vectors on a single A100 GPU with a batch size of 512 for 200 epochs. The training takes around 2 hours and only uses up to around 10 GB of GPU

Table 3.1: The effectiveness of sparsified models (D2S) and baselines. (${}^{\dagger} p < 0.05$ with paired two-tailed t-test comparing D2S to the dense model with Bonferroni correction)

Model	MSCOCO (5k)				Flickr30k (1k)			
	R@1 \uparrow	R@5 \uparrow	MRR@10 \uparrow	FLOPs \downarrow	R@1 \uparrow	R@5 \uparrow	MRR@10 \uparrow	FLOPs \downarrow
<i>T2I Dense Retrieval</i>								
COOKIE (Wen et al., 2021)	46.6	75.2	-	-	68.3	91.1	-	-
COTS (5.3M) (Lu et al., 2022)	50.5	77.6	-	-	75.2	93.6	-	-
ALBEF (Li et al., 2021a)	53.1	79.3	64.3	-	79.1	94.9	86.6	-
BLIP (Li et al., 2022b)	57.3	81.8	67.8	-	83.2	96.7	89.3	-
<i>T2I Sparse Retrieval</i>								
VisualSparta	45.1	73.0	-	-	57.1	82.6	-	-
STAIR (zero-shot)	41.1	56.4	-	-	66.6	88.7	-	-
LexLIP (4.3M)	51.9	78.3	-	-	76.7	93.7	-	-
LexLIP (14.3M)	53.2	79.1	-	-	78.4	94.6	-	-
D2S (ALBEF, $\eta = 1e - 3$)	49.6 †	77.7 †	61.4 †	18.7	74.2 †	93.8 †	82.6 †	21.7
D2S (ALBEF, $\eta = 1e - 5$)	50.7 †	78.2 †	62.4 †	74.2	75.4 †	94.3 †	83.6 †	64.3
D2S (BLIP, $\eta = 1e - 3$)	51.8 †	79.3 †	63.4 †	11.5	77.1 †	94.6 †	84.6 †	9.9
D2S (BLIP, $\eta = 1e - 5$)	54.5†	80.6†	65.6†	78.4	79.8†	95.9†	86.7†	39.5

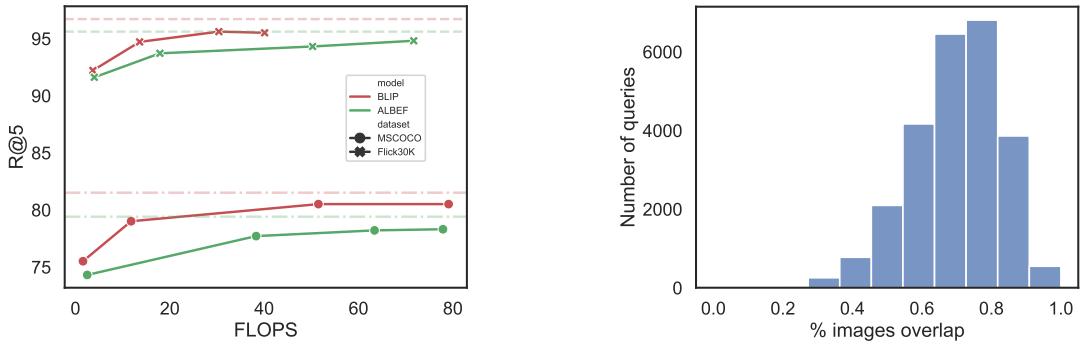
memory. We set the temperature τ to 0.001 and experiment with sparse regularization weights $\eta \in [1e - 5, 1e - 2]$.

3.5.2 Results and Discussion

RQ2.1: How effective is the proposed method for converting dense representations to sparse? We trained various Dense2Sparse models (D2S) using our proposed training method with different sparse regularization weights ranging from $1e - 5$ to $1e - 2$. Figure 3.2a illustrates the effectiveness and efficiency of these variations, with detailed results presented in Table 3.1. Firstly, we observe that increasing the sparse regularization weight enhances model efficiency (reduced FLOPs) but reduces its effectiveness (lower Recall and MRR). On the MSCOCO dataset, our most efficient sparse BLIP model ($\eta = 1e - 2$) achieves a R@1 of 47.2 and MRR@10 of 58.5 with the lowest FLOPs value of 1.6. Relaxing the regularization weight to $1e - 3$ results in an approximately 10% increase in R@1 to 51.8 and a similar rise in MRR@10 to 63.4, albeit at the expense of around 7 times higher FLOPs (less efficient).

Further relaxing the sparse regularization gradually brings the sparsified model’s effectiveness closer to the original dense model, while reducing the efficiency. The most effective sparsified BLIP model with $\eta = 1e - 5$ performs competitively with the original dense version (54.5 vs. 57.3) and outperforms other dense baselines.

Additionally, we observe a diminishing gap between dense and sparsified models



(a) Efficiency vs. effectiveness of sparsified models

(b) Fraction of overlapping images in the top-10 by sparsified and dense BLIP model.

Figure 3.2: Sparisified models compared to original dense models.

as we assess recalls at higher cutoff positions, such as $R@5$ and $R@10$. Similar trends are observed across different datasets, including Flickr30k and MSCOCO, as well as among different dense models, including BLIP and ALBEF. This indicates the broad applicability of our proposed approach to diverse datasets and models.

RQ2.2: *How does the proposed sparsified model compare to state-of-the-art multi-modal LSR models?* To answer this research question, we compare our sparsified models with existing LSR baselines, namely Visual Sparta, STAIR, and LexLIP. Currently, neither the code nor the checkpoints for these baselines are publicly available. Therefore, we rely on the numbers reported in their respective papers for comparison, excluding the FLOPs.

STAIR and LexLIP are two of the most recent multimodal LSR approaches, both trained on large datasets, with STAIR utilizing 1 billion internal text-image pairs. In contrast, our proposed method leverages pretrained dense retrieval models to efficiently learn a lightweight sparse projection for converting dense vectors to sparse vectors.

The effectiveness of our methods and the baselines on MS-COCO and Flickr30k is presented in Table 3.1. Notably, our efficient model, D2S(BLIP, $\eta = 1e - 3$), performs competitively with LexLIP trained on 4.3 million text-image pairs at $R@1$. Its $R@5$ is slightly better than LexLIP (4.3M) and comparable to the LexLIP model trained on 14.3 million pairs. With a lower sparse regularization, our D2S(BLIP, $\eta=1e-5$) model significantly outperforms all baselines on both MSCOCO and Flickr30k. On MSCOCO, its $R@1$ is 21%, 5%, and 2.8% higher than the $R@1$ of Visual Sparta, LexLIP (4.3M), and LexLIP (14.3M), respectively. All our models outperform Visual Sparta and STAIR, although this comparison with STAIR uses a zero-shot setting, because we lack access to their code and checkpoints for fine-tuning STAIR further with in-domain data.

We kept the dense encoders frozen, so the effectiveness of our sparsified models is inherently bounded by the dense results. Our sparsified ALBEF models, for exam-

Table 3.2: The dimension co-activation effect of Dense2Sparse (D2S) variations.

Model (D2S variations)	MSCOCO (5k)				Flickr30k (1k)			
	R@1↑	R@5↑	MRR@10↑	FLOPs↓	R@1↑	R@5↑	MRR@10↑	FLOPs↓
(BLIP, $\eta = 1e - 3$, $exp = 0$)	45.5	73.0	57.3	2.8	68.9	89.5	77.8	3.0
(BLIP, $\eta = 1e - 3$, $exp = 1$)	53.4	80.0	64.6	49.1	79.5	95.5	86.4	50.3
(BLIP, $\eta = 1e - 3$, $exp = c$)	51.9	79.0	63.4	11.8	77.3	94.7	84.8	13.6
(BLIP, $\eta = 1e - 3$, $exp = c + w$)	51.8	79.3	63.4	11.5	77.1	94.6	84.6	9.9
(BLIP, $\eta = 1e - 5$, $exp = 0$)	47.2	74.4	58.8	3.2	72.3	91.8	80.7	3.5
(BLIP, $\eta = 1e - 5$, $exp = 1$)	55.9	81.3	66.8	343	81.4	96.0	87.7	213
(BLIP, $\eta = 1e - 5$, $exp = c$)	54.7	80.5	65.8	79.1	79.9	95.5	86.7	40.1
(BLIP, $\eta = 1e - 5$, $exp = c + w$)	54.5	80.6	65.6	78.4	79.8	95.9	86.7	39.5
(ALBEF, $\eta = 1e - 3$, $exp = 0$)	43.8	71.8	55.7	2.5	65.8	88.3	75.4	3.0
(ALBEF, $\eta = 1e - 3$, $exp = 1$)	50.9	78.4	62.5	68.2	75.7	94.2	83.8	61.9
(ALBEF, $\eta = 1e - 3$, $exp = c$)	49.7	77.7	61.5	38.3	74.6	93.7	82.8	17.9
(ALBEF, $\eta = 1e - 3$, $exp = c + w$)	49.6	77.7	61.4	18.7	74.2	93.8	82.6	21.7
(ALBEF, $\eta = 1e - 5$, $exp = 0$)	45.9	73.9	83.0	3.4	68.1	90.0	77.6	3.2
(ALBEF, $\eta = 1e - 5$, $exp = 1$)	52.4	78.7	63.7	283	77.2	94.6	84.8	210
(ALBEF, $\eta = 1e - 5$, $exp = c$)	51.2	78.3	62.8	77.9	76.4	94.8	84.0	71.7
(ALBEF, $\eta = 1e - 5$, $exp = c + w$)	50.7	78.2	62.4	74.2	75.4	94.3	83.6	64.3

ple, exhibit slightly lower overall effectiveness since their corresponding dense performance is lower than that of BLIP’s dense scores. Nonetheless, our sparsified ALBEF models are also comparable with LexLIP variants.

RQ2.3: *How does the proposed training method impact dimension co-activation and semantic deviation?* As discussed in Section 3.3, high co-activation increases posting list length, impacting inverted index efficiency. We examine this impact by analyzing FLOPs alongside model effectiveness metrics. Table 3.1 presents results for models trained with our method and three baseline variants, with fixed expansion rates of 0 and 1 in the first two baselines. The third baseline ($exp = c$) explores the influence of word-level expansion control, excluding it from our training method.

At an expansion rate of zero, models project the caption’s dense vector only onto terms from the caption, with all other projections forced to zero. The image projector must then learn to align the image vector with terms in the paired captions. Conversely, setting exp to 1 gives the model the freedom to project onto any output vectors, making it more inclined toward dimension co-activation.

In Table 3.2, rows with ($exp = 0$) show models with no expansion, resulting in remarkably low FLOPs, with each query averaging 2 to 3 overlapping terms with each document. However, disabling expansion reduces the model’s ability for semantic matching, leading to modest effectiveness (45–47 R@1 on MSCOCO and 68–72 R@1 on Flickr30k with varying sparsity). Enabling non-regulated expansion ($exp = 1$)

Table 3.3: Semantic deviation on different Dense2Sparse (D2S) variations. ($\dagger p < 0.01$ with paired two-tailed t -test comparing $exp=c$ to $exp=1$)

Model (D2S variations)	MSCOCO (5k)		Flickr30k (1k)	
	Exact@20	Semantic@20	Exact@20	Semantic@20
(BLIP, $\eta = 1e - 5$, $exp = c$)	20.0 [†]	60.1 [†]	18.3 [†]	58.0 [†]
(BLIP, $\eta = 1e - 5$, $exp = 1$)	6.9	48.5	3.2	40.7
(BLIP, $\eta = 1e - 3$, $exp = c$)	25.0 [†]	63.2 [†]	23.1 [†]	60.6 [†]
(BLIP, $\eta = 1e - 3$, $exp = 1$)	2.5	42.0	2.2	41.1
(ALBEF, $\eta = 1e - 5$, $exp = c$)	20.5 [†]	61.0 [†]	19.2 [†]	59.8 [†]
(ALBEF, $\eta = 1e - 5$, $exp = 1$)	5.6	43.5	1.2	40.5
(ALBEF, $\eta = 1e - 3$, $exp = c$)	15.1 [†]	51.3 [†]	19.6 [†]	56.4 [†]
(ALBEF, $\eta = 1e - 3$, $exp = 1$)	1.6	40.6	1.3	41.5

significantly improves model effectiveness (50–55 R@1 on MSCOCO and 75–79 R@1 on Flickr30k with various regularization weights). However, this improvement comes at the cost of substantially increased FLOP scores, sometimes by up to 100 times, making sparsified vectors very computationally expensive. Ultimately, the resulting models behave like dense models, which is an undesired effect.

Our training method, which incorporates expansion control at the caption and word levels, is designed to gradually transition from one extreme ($exp = 0$) to the other ($exp = 1$). During training, we allow a likelihood of expansion, which increases progressively to over time. However, we also introduce random elements, represented by a random variable, to remind the model to remain faithful to the original captions/images.

The results, displayed in rows labeled with $exp = c + w$, demonstrate that our approach strikes a better balance between efficiency and effectiveness. It achieves competitive levels of effectiveness compared to models with $exp = 1$ while requiring only half or a third of the computational operations (FLOPs). For example, on MSCOCO with the BLIP model, Dense2Sparse ($\eta = 1e - 3$) achieves a performance of 51.8 R@1 (compared to 53.4 when $exp = 1$) with just 11.8 FLOPs, making it four times more efficient than the $exp = 1$ baseline. With the same setting, our method achieves 14% higher R@1 and 11% higher MRR@10 than the baseline with no expansion ($exp = 0$). Compared to the baseline without word-level expansion control, no significant differences are observed in terms of efficiency and effectiveness. Thus, caption-level expansion control alone seems sufficient for achieving reasonable efficiency and effectiveness. Similar results are noted across various settings, datasets, and dense models.

Sparse representations contain interpretable output dimensions aligned with a vo-

Table 3.4: Examples of semantic deviation. We show the top-10 terms per model.

Caption, Image	D2S ($\eta = 1e - 3$, $exp = c$)	D2S ($\eta = 1e - 3$, $exp = 1.0$)
A man with a red helmet dirt, mo, motor, motorcycle, , accent " yourself natural on a small moped on a dirt bike, red, riding, features, sol- may while officer english road dier, ##oot		dirt, mo, motor, motorcycle, , accent " yourself natural on a small moped on a dirt bike, red, riding, features, sol- may while officer english road dier, ##oot ac
	mountain mountains bike bee accent ship natural de dirt mo red path ##oot person crown yourself " ra now man riding bicycle	wild
A women smiling really big while holding a Wii remote.	lady woman smile women remote laughing wii smiling video controller	, kai called forces rush lee war oil like ##h
	smile after green woman smil- tall kai forces oil rush met ing sweater remote lady wii her war college thus there	
A couple of dogs sitting in the front seats of a car.	dogs dog car backseat seat couple vehicle sitting two puppy	, electric stood forest na- tional master help arts fc
	dog car dogs puppy out vehicle pup inside early open	stood forest national elec- tric master twice grant men para yet

cabulary. However, training a D2S model without our expansion regulation leads to semantic deviation, turning vocabulary terms into non-interpretable latent dimensions. We assess this effect using Exact@k and Semantic@k metrics (defined in Section 4.2), reporting results in Table 3.3 and providing qualitative examples in Table 3.4.

Uncontrolled models (with $exp = 1$) exhibit lower Exact@20 and Semantic@20 than our expansion-controlled models ($exp = c$). In the top 20 terms of uncontrolled models, only one or none are in the original captions, while controlled models generate 3 to 5 caption terms. The low Semantic@20 of the uncontrolled models also suggests low relatedness of output terms to the caption terms. This implication could be further supported by the examples demonstrated in Table 3.4. Uncontrolled models generate random terms, while our method produces terms that more faithfully reflect captions and images. Most top-10 terms from our method are relevant to the input, including a mix of original terms and synonyms (e.g., “dog” vs. “puppy”, “car” vs. “vehicle”).

RQ2.4: *Is the sparsified model faithful to the dense model?* This research question aims to analyze the faithfulness of sparsified models to their original dense models. We report in Table 3.5 the Pearson correlation calculated for various effectiveness metrics of dense

Table 3.5: Correlation between dense and different variations of Dense2Sparse (D2S).

Model (D2S variations)	MSCOCO (5k)			Flickr30k (1k)		
	$\rho\text{-R@1}\uparrow$	$\rho\text{-R@5}\uparrow$	$\rho\text{-MRR@10}\uparrow$	$\rho\text{-R@1}\uparrow$	$\rho\text{-R@5}\uparrow$	$\rho\text{-MRR@10}\uparrow$
(BLIP, $\eta = 1e - 2$)	61.0	65.7	72.3	54.7	55.0	63.9
(BLIP, $\eta = 1e - 3$)	74.0	76.9	83.8	66.2	65.5	73.6
(BLIP, $\eta = 1e - 4$)	79.7	82.1	88.2	71.6	72.8	79.3
(BLIP, $\eta = 1e - 5$)	81.2	83.8	89.2	74.3	74.0	81.1
(ALBEF, $\eta = 1e - 2$)	64.4	68.7	75.5	57.7	57.0	67.5
(ALBEF, $\eta = 1e - 3$)	73.1	76.7	83.5	68.8	69.0	77.2
(ALBEF, $\eta = 1e - 4$)	78.1	80.7	87.2	73.2	74.6	81.3
(ALBEF, $\eta = 1e - 5$)	78.2	81.3	87.3	74.2	72.5	82.0

and sparsified queries. The results show that the correlation between sparsified and dense models is consistently positive and tends to increase as we relax the sparse regularization. Furthermore, as we consider higher cutoff values (R@1, R@5, MRR@10), the correlation tends to increase as the performance gap between the two systems narrows. Manually comparing the top-10 ranked images of the most differing queries, we find that while the two models rank top-10 images differently, there are a lot of common images (including the golden image) that look equally relevant to the query. Figure 3.2b shows that a high ratio (average: 70%) of the top-10 images appear in both dense and sparse ranking lists. This analysis shows that the sparsified model is reasonably faithful to the dense model, suggesting that the sparse output terms could potentially be used for studying the semantics of dense vectors.

3.5.3 Retrieval Latency of Dense and Sparsified Models

We discussed the average FLOPs of sparsified models for retrieval efficiency. We now present query throughput and retrieval latency results in Table 3.6. Using Faiss (Johnson et al., 2019) and PISA (Mallia et al., 2019; MacAvaney and Macdonald, 2022) on a single-threaded AMD Genoa 9654 CPU, the dense BLIP model with Faiss HNSW is exceptionally fast, outperforming D2S models with PISA. D2S models with query expansion ($exp=c$) are slower due to high FLOPs and possibly LSR known limitations (Mackenzie et al., 2021). Removing expansion terms ($exp=0$) improves latency (FLOPs similar to DistilSPLADE (Formal et al., 2021; Formal et al., 2022)) but is still approximately 30× slower than dense retrieval. To balance efficiency and effectiveness of D2S, we propose using the inverted index with original query terms for retrieval, followed by re-scoring with expansion terms. With our simple iterative implementation, this approach proves effective, especially for retrieving fewer images per query. Surprisingly, indexing D2S models with Faiss HNSW competes well with PISA, par-

Table 3.6: Retrieval latency (CPU - 1 thread) of D2S models on 123k MSCOCO images.

Model	FLOPS	Throughput (q/s)			Latency (ms)		
		@10	@100	@1000	@10	@100	@1000
Dense (BLIP, HNSW, Faiss)	-	13277	9739	7447	0.08	0.10	0.14
D2S (BLIP, $\eta = 1e - 3$, exp=c, PISA)	11.5	6	5	5	156.60	183.42	193.46
D2S (BLIP, $\eta = 1e - 3$, exp=o, PISA)	2.8	449	284	160	2.23	3.52	6.25
No Expansion >> Expansion	-	369	120	18	2.70	8.31	54.05
D2S (BLIP, $\eta = 1e - 5$, exp=c, PISA)	78.4	<1	<1	<1	>300	>600	>700
D2S (BLIP, $\eta = 1e - 5$, exp=o, PISA)	3.2	230	146	90	4.34	6.85	11.04
No Expansion >> Expansion	-	189	70	11	5.30	14.37	86.66
D2S (BLIP, HNSW, Faiss)	-	262	262	256	3.82	3.82	3.90

ticularly at higher cut-off values (100, 1000).

3.6 CONCLUSION

We have focused on the problem of efficiently transforming a pretrained dense retrieval model into a sparse model. We show that training a projection layer on top of dense vectors with the standard contrastive learning technique leads to the problems of dimension co-activation and semantic deviation. To mitigate these issues, we propose a training algorithm that uses a Bernoulli random variable to control the term expansion. Our experiments show that our Dense2Sparse sparsified model trained with the proposed algorithm suffers less from those issues. In addition, our sparsified models perform competitively to the state-of-the-art multi-modal LSR, while being faithful to the original dense models.

Consequently, we conclude for RQ2 that in the vision-language domain, learned sparse retrieval techniques can be applied by converting dense representations into sparse ones, showing promising results in both effectiveness and efficiency.

REPRODUCIBILITY

To ensure the reproducibility of the findings presented in this chapter, we have made our code publicly accessible at <https://github.com/thongnt99/lsr-multimodal>.

4

SHORTCUTS IN VISION-LANGUAGE REPRESENTATION LEARNING

We continue our investigation by focusing on the problem of shortcut learning in the context of contrastive vision-language (VL) representation learning with multiple captions per image. We assume that all captions associated with the image contain both shared and caption-specific information.

Specifically, we want to investigate if in such cases model learns all the information available in the captions or if it learns a shortcut, i.e., a subset of information that minimizes the loss but is not necessarily useful for the task at hand.

Hence, we ask the following research question:

RQ3: In the context of vision-language representation learning with multiple captions per image, to what extent does the presence of a shortcut hinder learning task-optimal representations?

To address this question, we propose and develop the framework for synthetic shortcuts for vision-language (SVL). This framework allows us to augment image-caption tuples with additional identifiers that do not bear any semantic meaning and therefore study the problem of shortcut learning in a controlled way.

We experiment with the two following distinct models: CLIP, a large-scale model that we fine-tune, and VSE++, a smaller model trained from scratch. We show that contrastive VL methods tend to depend on shortcuts and suppress task-relevant information.

This chapter was published in the Transactions on Machine Learning Research (TMLR 2024) under the title “Demonstrating and Reducing Shortcuts in Vision-Language Representation Learning” (Bleeker et al., 2024).

4.1 INTRODUCTION

Recent work on understanding the internal mechanisms of representation learning has brought to attention the problem of shortcut learning (Robinson et al., 2021; Chen et al., 2021; Scimeca et al., 2022). While there are multiple definitions of shortcut learning (e.g., Geirhos et al., 2020; Wiles et al., 2022), in this chapter we define *shortcuts* as *easy-to-learn discriminatory features that minimize the (contrastive) optimization objective but are not necessarily sufficient for solving the evaluation task*. More specifically, we focus on the problem of shortcut learning in the relatively unexplored context of VL representation learning with multiple matching captions per image.

Contrastive learning (CL) plays a crucial role in VL representation learning. Despite the success of non-contrastive approaches, e.g., (Bardes et al., 2022), the dominant paradigm in VL representation learning revolves around either fully contrastive strategies (Faghri et al., 2018; Li et al., 2019a; Jia et al., 2021; Radford et al., 2021) or a combination of contrastive methods with additional objectives (Li et al., 2021a; Zeng et al., 2022; Li et al., 2022b; Li et al., 2023a). It is standard practice in contrastive VL representation learning to sample batches of image-caption pairs and maximize the alignment between the representations of the matching images and captions (Radford et al., 2019; Jia et al., 2021). Given that the typical VL benchmarks, e.g., Flickr30k (Young et al., 2014) and MS-COCO Captions (Lin et al., 2014; Chen et al., 2015), are constructed in such a way that each image is associated with multiple captions, each caption can be seen as a different *view* of the image it describes. Therefore, CL with multiple captions per image can be seen as CL with multiple views, where each caption provides a different view of the scene depicted in the image.

CL with multiple views, where each view represents a different observation of the same datapoint, has proven to be effective for general-purpose representation learning (Hjelm et al., 2019; Chen et al., 2020a; Tian et al., 2020a). The goal of multi-view (contrastive) representation learning methods is to learn representations that remain invariant to a shift of view, which is achieved by maximizing alignment between embeddings of similar views. A core assumption within the multi-view representation learning literature is that task-relevant information is shared across views whereas task-irrelevant information is not shared, given a downstream evaluation task (Zhao et al., 2017; Federici et al., 2020; Tian et al., 2020a; Shwartz-Ziv and LeCun, 2023).

An open challenge in the multi-view representation learning domain concerns *learning representations that contain task-relevant information that is not shared among different views, i.e., that may be unique for some views* (Shwartz-Ziv and LeCun, 2023; Zong et al., 2023). In the case of image-caption datasets where each image is paired with at least one corresponding caption, the captions matching the same image do not necessarily share the same information as each caption is distinct and may describe different

aspects of the image (Biten et al., 2022).

Figure 4.1 illustrates the concept of shared vs. caption-specific task-relevant information. The image is accompanied by two captions: ‘a couple of boats and a red car’ (x_{C_A}) and ‘a couple of boats and a car on a street’ (x_{C_B}). The shared information between the captions includes ‘couple of boats’ and ‘car’. Caption x_{C_A} provides unique information by describing the car as ‘red’. Caption x_{C_B} adds unique contextual details about the location with the phrase ‘on a street’. To learn task-optimal representations, it is essential to integrate both the shared and unique information from these captions. Furthermore, given the typical quality of captions of image-caption datasets (Chen et al., 2015), we assume that all information present in the captions is relevant. Hence, each image-caption pair may contain both *shared* task-relevant information, i.e., information shared across all the captions in the tuple, and *unique* task-relevant information, i.e., information not shared with other captions. Therefore, learning task-optimal representations for the image implies learning all task-relevant information that comprises both shared and caption-specific information.

Another problem of CL approaches is related to *feature suppression*. (Shwartz-Ziv and LeCun, 2023) argue that although contrastive loss functions lack explicit information-theoretical constraints aimed at suppressing non-shared information among views, the learning algorithm benefits from simplifying representations by suppressing features from the input data that are not relevant for minimizing the contrastive loss. Furthermore, Robinson et al. (2021) demonstrate that contrastive loss functions are susceptible to solutions that suppress features from the input data. In the case of VL, CL with multiple captions per image where at least one caption contains caption-specific information, the image representation can never have a perfect alignment with all matching captions. This is due to the misalignment that happens when encoding unique information for the other captions. Therefore, it is unclear whether contrastive methods can learn task-optimal representations, i.e., representations that contain all information present in the captions associated with the image, or if they learn only the minimal shared information, i.e., information shared between the image and all captions that are sufficient to minimize the contrastive discrimination objective. An

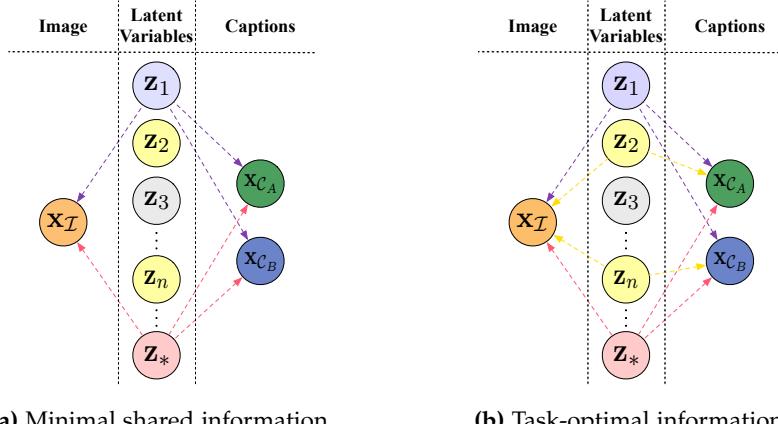


$\textcolor{purple}{x}_{C_A}$: a couple of boats and a red car

$\textcolor{blue}{x}_{C_B}$: a couple of boats and car on a street

Figure 4.1: Shared vs. caption-specific information given an example of one image and two associated captions x_{C_A} and x_{C_B} . The purple color indicates information shared between the image and both captions. The green color indicates task-relevant information specific for x_{C_A} . The blue color indicates task-relevant information specific for x_{C_B} .

illustration of minimal shared information and a task-optimal representation is given in Figure 4.2.



(a) Minimal shared information.

(b) Task-optimal information.

Figure 4.2: Synthetic shortcuts in the context of minimal shared and task-optimal information for vision-language representation learning with multiple captions per image. The purple color represents features shared among the image and all captions (minimal shared information). The yellow color represents caption-specific features (unique information). The grey color indicates features that are not present in both the image and any of the captions (task-irrelevant information). The red color indicates synthetic shortcuts. We demonstrate that while shortcuts exist in both scenarios, minimal shared information also includes information shared among the image and all associated captions, whereas task-optimal information combines both minimal shared information and caption-specific information.

Motivated by the abovementioned problems, we address the following question:

In the context of VL representation learning with multiple captions per image, to what extent does the presence of a shortcut hinder learning task-optimal representations?

To answer this question, we investigate the problem of shortcut learning for VL representation learning with multiple captions per image. We do this by introducing the framework for *synthetic shortcuts for vision-language* (SVL) for adding additional, easily identifiable information to image-caption tuples. The information that we add is represented as identifiers that are applied to both image and caption; these identifiers do not bear any semantic meaning. The identifiers provide additional shared information between the image and captions, which is a subset of the total shared information between the image and the caption. For details and examples of shortcuts, refer to Section 4.3, where Figure 4.4 illustrates an example of an image-caption pair with a shortcut added. The framework allows us to investigate how much the encoder model relies on the added shortcut during training and evaluation, and hence how much of the relevant information is still captured if a shortcut solution is available. Overall, our SVL framework allows us to investigate the shortcut learning problem in a controlled way. We focus on image-caption retrieval (ICR) as an evaluation task because contrastive losses directly optimize for the ICR evaluation task, which assesses

the quality of the learned representations by computing a similarity score between images and captions (Radford et al., 2021; Yuksekgonul et al., 2023). To investigate the problem, we run experiments on two distinct models: (i) CLIP (Radford et al., 2019), a large-scale model that we fine-tune; and (ii) VSE++ (Faghri et al., 2018), a relatively small model that we train from scratch. We evaluate the models’ performance on the Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014; Chen et al., 2015) and benchmarks. The benchmarks are constructed in such a way that each image is associated with five captions and each caption represents a concise summary of the corresponding image.

Therefore, the contributions of this chapter are two-fold:

I We present a framework for investigating the problem of shortcut learning for contrastive vision-language representation learning in a controlled way: We introduce the framework for *synthetic shortcuts for vision-language*. The framework enables the injection of synthetic shortcuts into image-caption tuples in the training dataset. We use the framework to investigate and understand the extent to which contrastive VL models rely on shortcuts when a shortcut solution is available. We run our experiments using CLIP and VSE++, two distinct vision-language models (VLMs). We evaluate the models’ performance on the Flickr30k and MS-COCO benchmarks. We evaluate the effectiveness of contrastive VL models by comparing their performance with and without synthetic shortcuts. We demonstrate that both models trained from scratch and fine-tuned, large-scale pre-trained foundation models mainly rely on shortcut features and do not learn task-optimal representations. Consequently, we show that contrastive losses mainly capture the easy-to-learn discriminatory features that are shared among the image and all matching captions, while suppressing other task-relevant information. Hence, we argue that contrastive losses are not sufficient to learn task-optimal representations for VL representation learning.

II We evaluate two shortcut learning reduction methods on our proposed training and evaluation framework: We investigate latent target decoding (LTD) and implicit feature modification (IFM) using our SVL training and evaluation framework. While both methods improve performance on the evaluation task, our framework poses challenges that existing shortcut reduction techniques can only partially address, as the performance is not on par with models trained without synthetic shortcuts. These findings underline the importance and complexity of our framework in studying and evaluating shortcut learning within the context of contrastive VL representation learning.

4.2 BACKGROUND AND ANALYSIS

In this section, we present the notation, setup, and assumptions on which we base the chapter. Additionally, we conduct an analysis of contrastive VL representation learning with multiple captions per image.

4.2.1 Preliminaries

Notation. We closely follow the notation from (Brown et al., 2020; Hendriksen et al., 2023; Bleeker et al., 2022). Let \mathcal{D} be a dataset of N image-caption tuples: $\mathcal{D} = \left\{ \left(\mathbf{x}_{\mathcal{I}}^i, \{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k \right) \right\}_{i=1}^N$. Each tuple $i \in N$ contains one image $\mathbf{x}_{\mathcal{I}}^i$ and k captions $\mathbf{x}_{\mathcal{C}_j}^i$, where $1 \leq j \leq k$. All captions in tuple $i \in N$ are considered as matching captions w.r.t. image $\mathbf{x}_{\mathcal{I}}$ in the tuple i . The latent representation of an image-caption pair from a tuple i is denoted as $\mathbf{z}_{\mathcal{I}}^i$ and $\mathbf{z}_{\mathcal{C}_j}^i$, respectively. During training, we sample image-caption pairs from the dataset \mathcal{D} and optimize for the evaluation task T . We include all captions in the dataset once per training epoch, hence, each image is sampled k times.

Given an image $\mathbf{x}_{\mathcal{I}}$, a set of k associated captions $K = \{\mathbf{x}_{\mathcal{C}_j}\}_{j=1}^k$, and one caption randomly sampled from the set $\mathbf{x}_{\mathcal{C}} \in K$, we define the following representations: (i) $\mathbf{z}_{\mathcal{C} \rightarrow \mathcal{I}}^{SUF}$ as *sufficient* representation of the caption $\mathbf{x}_{\mathcal{C}}$ that describes the image $\mathbf{x}_{\mathcal{I}}$; (ii) $\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{SUF}$ as representation of the image $\mathbf{x}_{\mathcal{I}}$ *sufficient for the caption* $\mathbf{x}_{\mathcal{C}}$; (iii) $\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{MIN}$ as representation of the image $\mathbf{x}_{\mathcal{I}}$ that is *minimally sufficient for the caption* $\mathbf{x}_{\mathcal{C}}$; and (iv) $\mathbf{z}_{\mathcal{I} \rightarrow K}^{OPT}$ as representation of the image $\mathbf{x}_{\mathcal{I}}$ that is *optimal for the set of captions* K given the task T .

In addition, we write S_{SynSC} for a synthetic shortcut, S for the original shared information, i.e., information that does not contain synthetic shortcuts, S^+ for the shared information that includes a synthetic shortcut, and R^+ for task-relevant information that contains a synthetic shortcut.

In the context of task relevance, we define R and $\neg R$ as task-relevant and task-irrelevant information, respectively, and C as task-relevant information specific for caption $\mathbf{x}_{\mathcal{C}}$. See Appendix 4.A, Table 4.A.1 for the notation overview.

Setup. We work with a dual-encoder setup, with an image encoder and a caption encoder that do not share parameters. The *image encoder* $f_{\theta}(\cdot)$ takes an image $\mathbf{x}_{\mathcal{I}}$ as input and returns its latent representation: $\mathbf{z}_{\mathcal{I}} := f_{\theta}(\mathbf{x}_{\mathcal{I}})$. Similarly, the *caption encoder* $g_{\phi}(\cdot)$ takes a caption $\mathbf{x}_{\mathcal{C}}$ as input, and encodes the caption into a latent representation: $\mathbf{z}_{\mathcal{C}} := g_{\phi}(\mathbf{x}_{\mathcal{C}})$. Both $\mathbf{z}_{\mathcal{C}}$ and $\mathbf{z}_{\mathcal{I}}$ are unit vectors projected into d -dimensional multi-modal space: $\mathbf{z}_{\mathcal{C}} \in \mathcal{R}^d$, $\mathbf{z}_{\mathcal{I}} \in \mathcal{R}^d$. For an overview of notation, we refer to Appendix 4.A, Table 4.A.1.

Assumptions. Given an image-caption tuple, we assume that each caption in the tuple is distinct from the other captions in the tuple. We also assume that each caption in

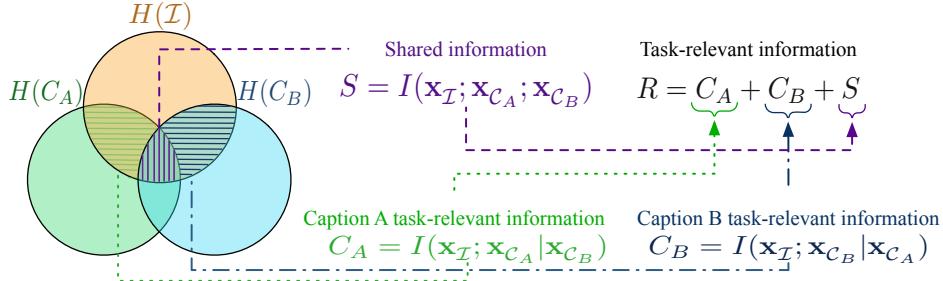


Figure 4.3: We define $H(\mathbf{x}_I)$ as image information, $H(\mathbf{x}_{C_A})$ and $H(\mathbf{x}_{C_B})$ as caption information; both captions only describe the information depicted in the image and contain shared and caption-specific information. We further define $C_A = I(\mathbf{x}_I; \mathbf{x}_{C_A} | \mathbf{x}_{C_B})$ and $C_B = I(\mathbf{x}_I; \mathbf{x}_{C_B} | \mathbf{x}_{C_A})$ as caption-specific information; $S = I(\mathbf{x}_I; \mathbf{x}_{C_A}; \mathbf{x}_{C_B})$ as shared information; $\neg R = H(\mathbf{x}_I | \mathbf{x}_{C_A}, \mathbf{x}_{C_B})$ as task-irrelevant information; $R = C_A + C_B + S$ as task-relevant information.

the tuple contains two types of task-relevant information: (i) shared information, i.e., information shared with other captions in the same tuple, and (ii) caption-specific information, i.e., information that is not shared with the other captions. For simplicity, we base our subsequent analysis on tuples where one image \mathbf{x}_I is associated with two captions \mathbf{x}_{C_A} and \mathbf{x}_{C_B} : $(\mathbf{x}_I, \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\})$. However, the analysis described in this section can be extended to a case with more than two captions. We treat images and captions as views and define \mathbf{x}_I , \mathbf{x}_{C_A} , and \mathbf{x}_{C_B} to be random variables of an image and two matching captions, with the joint distribution $p(\mathbf{x}_I, \mathbf{x}_{C_A}, \mathbf{x}_{C_B})$. For more details on assumptions and problem definition, we refer to Appendix 4.B.

4.2.2 Analysis of Contrastive Vision-Language Representation Learning for Multiple Captions per Image

InfoMax. We start our analysis of contrastive VL representation learning by introducing the InfoMax optimization objective, a typical loss for VL representation learning. The goal of an InfoMax optimization objective, e.g., InfoNCE (Oord et al., 2018), is to maximize the mutual information (MI) between the latent representations of two views of the same data (Tschanne et al., 2020). Therefore, the optimization objective is equivalent to: $\max_{f_\theta, g_\phi} I(\mathbf{z}_I; \mathbf{z}_C)$ where $\mathbf{z}_I = f_\theta(\mathbf{x}_I)$ and $\mathbf{z}_C := g_\phi(\mathbf{x}_C)$.

Minimally Sufficient Image Representation. During training, we sample batches of image-caption. The optimization involves maximizing the MI between the image representation \mathbf{z}_I and the matching caption representation \mathbf{z}_C . (Wang et al., 2022a) argue that, since all supervision information for one view (i.e., the image) comes from the other view (i.e., the caption), the representations learned contrastively are approximately minimally sufficient. Following (Tian et al., 2020b; Wang et al., 2022a), we extend the definition of sufficient representation to VL context and define sufficient caption representations, sufficient image representations, and minimally sufficient im-

age representation.

Definition 3 (Sufficient caption representation). *Given an image \mathbf{x}_I , and a set of matching captions $\mathcal{C} = \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\}$, the representation $\mathbf{z}_{\mathcal{C} \rightarrow I}^{SUF}$ of caption $\mathbf{x}_C \in \mathcal{C}$ is sufficient for image \mathbf{x}_I if, and only if, $I(\mathbf{z}_{\mathcal{C} \rightarrow I}^{SUF}; \mathbf{x}_I) = I(\mathbf{x}_C; \mathbf{x}_I)$.*

The sufficient caption representation $\mathbf{z}_{\mathcal{C} \rightarrow I}^{SUF}$ contains all the information about image \mathbf{x}_I in caption \mathbf{x}_C .

Definition 4 (Sufficient image representation). *Given an image \mathbf{x}_I , and a set of matching captions $\mathcal{C} = \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\}$, the representation $\mathbf{z}_{I \rightarrow \mathcal{C}}^{SUF}$ of image \mathbf{x}_I is sufficient for caption $\mathbf{x}_C \in \mathcal{C}$ if, and only if, $I(\mathbf{z}_{I \rightarrow \mathcal{C}}^{SUF}; \mathbf{x}_C) = I(\mathbf{x}_I; \mathbf{x}_C)$.*

Similarly, the sufficient image representation $\mathbf{z}_{I \rightarrow \mathcal{C}}^{SUF}$ contains all the shared information between an image \mathbf{x}_I and a caption \mathbf{x}_C . Note that a sufficient image representation can be sufficient w.r.t. multiple captions.

Definition 5 (Minimally sufficient image representation). *Given an image \mathbf{x}_I , and a set of matching captions $\mathcal{C} = \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\}$, the sufficient image representation $\mathbf{z}_{I \rightarrow \mathcal{C}}^{MIN}$ of image \mathbf{x}_I is minimally sufficient for caption $\mathbf{x}_C \in \mathcal{C}$ if, and only if, $I(\mathbf{z}_{I \rightarrow \mathcal{C}}^{MIN}; \mathbf{x}_I) \leq I(\mathbf{z}_{I \rightarrow \mathcal{C}}^{SUF}; \mathbf{x}_I)$, for all $\mathbf{z}_{I \rightarrow \mathcal{C}}^{SUF}$ that are sufficient.*

Intuitively, $\mathbf{z}_{I \rightarrow \mathcal{C}}^{MIN}$ comprises the smallest amount of information about \mathbf{x}_I (while still being sufficient) and, therefore, only contains the information that is shared with caption \mathbf{x}_C , i.e., the non-shared information is suppressed.

Task-Optimal Image Representation. The definition of task-optimal image representation is based on the notion of task-relevant information. In the context of VL representation learning with multiple captions per image, we define task-relevant information as all information described by the matching captions. That includes both caption-specific and shared information. Consequently, task-optimal image representation is image representation that is sufficient w.r.t. all matching captions.

Formally, following assumptions from Appendix 4.B.2, we define task-relevant information R as all the information described by the matching captions. The task-relevant information can be expressed as follows:

$$\begin{aligned} \underbrace{R}_{\text{Task-relevant information}} &= \underbrace{H(\mathbf{x}_I)}_{\text{Image information}} - \underbrace{H(\mathbf{x}_I \mid \mathbf{x}_{C_A}, \mathbf{x}_{C_B})}_{\text{Task-irrelevant information}} \\ &= \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_A} \mid \mathbf{x}_{C_B})}_{C_A\text{-specific task-relevant information}} + \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_B} \mid \mathbf{x}_{C_A})}_{C_B\text{-specific task-relevant information}} + \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_A}, \mathbf{x}_{C_B})}_{\text{Shared information}}. \end{aligned} \quad (4.1)$$

Similarly, task-irrelevant information $\neg R$ is the image information not described by the captions. Figure 4.3 illustrates both definitions.

The multi-view assumption states that task-relevant information for downstream tasks comes from the information shared between views (Shwartz-Ziv and LeCun, 2023). However, in the case of VL representation learning with multiple captions per image, task-relevant information R includes both shared information S , and caption-specific information C_A and C_B (Eq. 4.1).

Definition 6 (Task-optimal image representation). *Given an image \mathbf{x}_I , and a set of matching captions $\mathcal{C} = \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\}$, the representation $\mathbf{z}_{I \rightarrow C}^{OPT}$ is task-optimal image representation for all matching captions if, and only if, $I(\mathbf{z}_{I \rightarrow C}^{OPT}; \mathbf{x}_C) = I(\mathbf{x}_I; \mathbf{x}_C)$, for all $\mathbf{x}_C \in \mathcal{C}$.*

In other words, task-optimal image representations contain all the information that the image shares with the matching captions. Hence, a task-optimal image representation is sufficient w.r.t. all matching captions. The information contained in the task-optimal image representation includes both shared and caption-specific information. Therefore, a task-optimal image representation can never be a minimally sufficient image representation w.r.t. to a specific caption.

Theorem 1 (Suboptimality of contrastive learning with multiple captions per image). *Given an image \mathbf{x}_I , a set of matching captions $\mathcal{C} = \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\}$, and a contrastive learning loss function $\mathcal{L}_{\text{InfoNCE}}$ that optimizes for task T , image representations learned during contrastive learning will be minimally sufficient and will never be task-optimal image representations.*

The proof is provided in Appendix 4.C. Rephrasing Theorem 1, given an image and two captions that form two image-caption pairs, $(\mathbf{x}_I, \mathbf{x}_{C_A})$ and $(\mathbf{x}_I, \mathbf{x}_{C_B})$, and assuming that contrastive loss optimizes the image encoder to be minimally sufficient w.r.t. to caption \mathbf{x}_{C_A} during a training step, all task-relevant information C_B specific to caption \mathbf{x}_{C_B} will be suppressed in \mathbf{z}_I . Hence, the resulting image representation will not be optimal for the task T .

Theorem 1 highlights a gap between minimal sufficient representations learned during contrastive training with the InfoNCE loss and the task-optimal image representations in the context of learning VL representations with multiple captions per image. Although the InfoMax loss does not have an explicit constraint to compress information, prior work indicates that feature suppression is happening (Shwartz-Ziv and LeCun, 2023; Robinson et al., 2021). Hence, we question if contrastive loss can be used to learn task-optimal image representations in the context of multiple captions per image.

Furthermore, Theorem 1 implies that in the context of contrastive VL representation learning with multiple captions per image, the minimally sufficient representation, which discards non-shared information, is not the same as the task-optimal representation that comprises both caption-specific and shared information. This suggests that the features learned during contrastive learning might be shortcuts, i.e., easy-to-detect

discriminatory features that minimize the contrastive optimization objective but are not necessarily sufficient for solving the evaluation task. To examine this problem, we introduce a synthetic shortcuts framework that allows us to investigate the problem of suboptimality of contrastive learning with multiple captions per image in a controlled way.

4.3 SYNTHETIC SHORTCUTS TO CONTROL SHARED INFORMATION

In Section 4.2 we show the suboptimality of the contrastive InfoNCE loss with multiple captions per image. In the case of real-world VL datasets with multiple captions per image, there are no annotations that indicate the information shared between the image and captions and the information specific to each caption. Hence, we cannot directly measure how much of the shared and unique information is captured by the representations.

Synthetic Shortcuts. In this section, we introduce the training and evaluation framework for *synthetic shortcuts for vision-language (SVL)*. We denote the *synthetic shortcuts for image-caption data* as S_{SynSC} . The purpose of the framework is to introduce additional and easily identifiable information shared between an image and the matching captions that lacks any semantic meaning. The shortcuts we use in this chapter are represented as numbers that we add to images and captions. For images, we add the shortcut number by adding MNIST images as an overlay to the original images. For captions, we append the numbers of the shortcut as extra tokens at the end of the caption.

Figure 4.4 illustrates an example of an image-caption pair with an added shortcut. The example contains an image with the caption: ‘A player up to bat in a baseball game. 1 0 1 9 9 2.’ Here, ‘1 0 1 9 9 2’ is a shortcut added to both the image and the caption. For the image modality, we add the shortcut by overlaying MNIST images at the top of the original image. For the text modality, we append the shortcut as additional tokens at the end of the caption. This identifier provides an additional link between the image and the caption without carrying any semantic meaning related to their content. Additional examples are shown in Figure 4.D.1 in the Appendix 4.D.

If contrastive losses learn task-optimal representations, then the presence of synthetic shortcuts should not negatively impact the evaluation performance, since syn-



A player up to bat in a baseball game. 1 0 1 9 9 2

Figure 4.4: An image-caption pair from the MS-COCO dataset with a shortcut added to both the image and the caption.

thetic shortcuts represent additional information and the remaining task-relevant information is intact. By incorporating synthetic shortcuts into the image-caption dataset, the shared information would include the information that was originally shared and the synthetic shortcut: $S^+ = S + S_{SynSC}$. Hence, the task-relevant information would comprise caption-specific information that was originally shared and a synthetic shortcut: $R^+ = C_A + C_B + S + S_{SynSC}$. If injecting a synthetic shortcut influences the performance negatively, we can conclude that by learning to represent a synthetic shortcut the model suppresses other task-relevant information in favor of the shortcut, hence the representation is not task-optimal. The setup is inspired by the “datasets with explicit and controllable competing features,” introduced by (Chen et al., 2021), but we adapt this setup to the VL scenario.

For experiments, we use the Flickr30k and MS-COCO image-caption datasets, that consist of image-caption tuples, each image is associated with five captions. During training, we sample a batch \mathcal{B} of image-caption pairs $\mathcal{B} = \{(\mathbf{x}_{\mathcal{I}}^i, \mathbf{x}_{\mathcal{C}_j}^i), \dots\}_{i=1}^{|\mathcal{B}|}$, from dataset \mathcal{D} , and apply shortcut sampling. We inject the shortcuts in a manner that preserves the original information of the images and captions. Furthermore, we append the shortcut after applying data augmentations to ensure that the shortcut is present in both the images and captions (i.e., the shortcut is not augmented away). We refer to Figure 4.D.1 in the Appendix 4.D.4 for some examples. The training, evaluation, and implementation details of the shortcut sampling are provided in Appendix 4.D.4.

We define the following experimental setups:

I *No shortcuts*: As a baseline, we fine-tune a pre-trained CLIP (Radford et al., 2021) and train VSE++ (Faghri et al., 2018) from scratch on Flickr30k and MS-COCO, without using any shortcuts. The experimental setup for training both models is provided in Appendix 4.D.2 and 4.D.3. The goal of this setup is to show the retrieval evaluation performance without adding any shortcuts for both a large-scale pre-trained foundation model and a small-scale model trained from scratch.

II *Unique shortcuts*: We add a unique shortcut to each image-caption tuple $i \in \mathcal{D}$ in the dataset. In this setup, each image caption pair can be uniquely matched during training by only detecting the shortcut. For each tuple $i \in \mathcal{D}$, we use the number i as the number of the shortcut we inject to the image and captions in the tuple. If the contrastive loss learns task-optimal representations, the downstream evaluation performance should not decrease when training with unique shortcuts.

III *Unique shortcuts on only one modality*: To show that the shortcuts do not interfere with the original task-relevant information (S, C_A , and C_B) of the images and captions, we create a dataset with only shortcuts on either the image or caption

modality. Therefore, the shortcut cannot be used by the encoders to match an image-caption pair. Hence, we expect the encoders to ignore the shortcuts and extract the features from the original data similar to the features learned by the baseline models in experimental setup I.

IV N bits of shortcuts: In this setup, for each image-caption pair in the training batch \mathcal{B} , we randomly sample a shortcut number from the range $[0, 2^n]$, where n is the number of bits. The higher the value of n , the more image-caption pairs in the training batch will have by expectation a unique shortcut, and, the less the model has to rely on S and the remaining task-relevant information to solve the contrastive objective. The goal of this setup is to show that, the more unique (shortcut) information is present per sample in the batch, the less contrastive models rely on the remaining task-relevant information.

It should be noted that the shortcuts we add are independent of the image-caption pairs. However, the goal of the SVL framework is to measure the effect of the presence of additional easy-to-detect shared information on the learned representations.

Evaluation Method. To show the effect of the injected shortcuts on retrieval evaluation performance, we evaluate both with and without adding the shortcuts during evaluation. When training with unique shortcuts, we add a unique shortcut to each tuple in the test set as well. When training with shortcuts on either one of the two modalities, we only evaluate without shortcuts to show that training with shortcuts on one modality does not influence performance. When training with n bits of shortcuts, we add the shortcut $\text{mod}(i, n)$ (modulo) to each tuple i in the evaluation set, to make sure we use the same number of shortcuts during evaluation as during training.

4.4 SYNTHETIC SHORTCUTS AND THEIR IMPACT ON LEARNED REPRESENTATIONS

First, we train and evaluate both a CLIP and VSE++ without shortcuts on the Flickr30k and MS-COCO dataset for the image-caption retrieval task as a baseline. We use the recall sum (i.e., the sum of $R@1$, $R@5$, and $R@10$ for both image-to-text (i2t) and text-to-image (t2i) retrieval) as evaluation metric (see Appendix 4.B.1 for the evaluation task description).

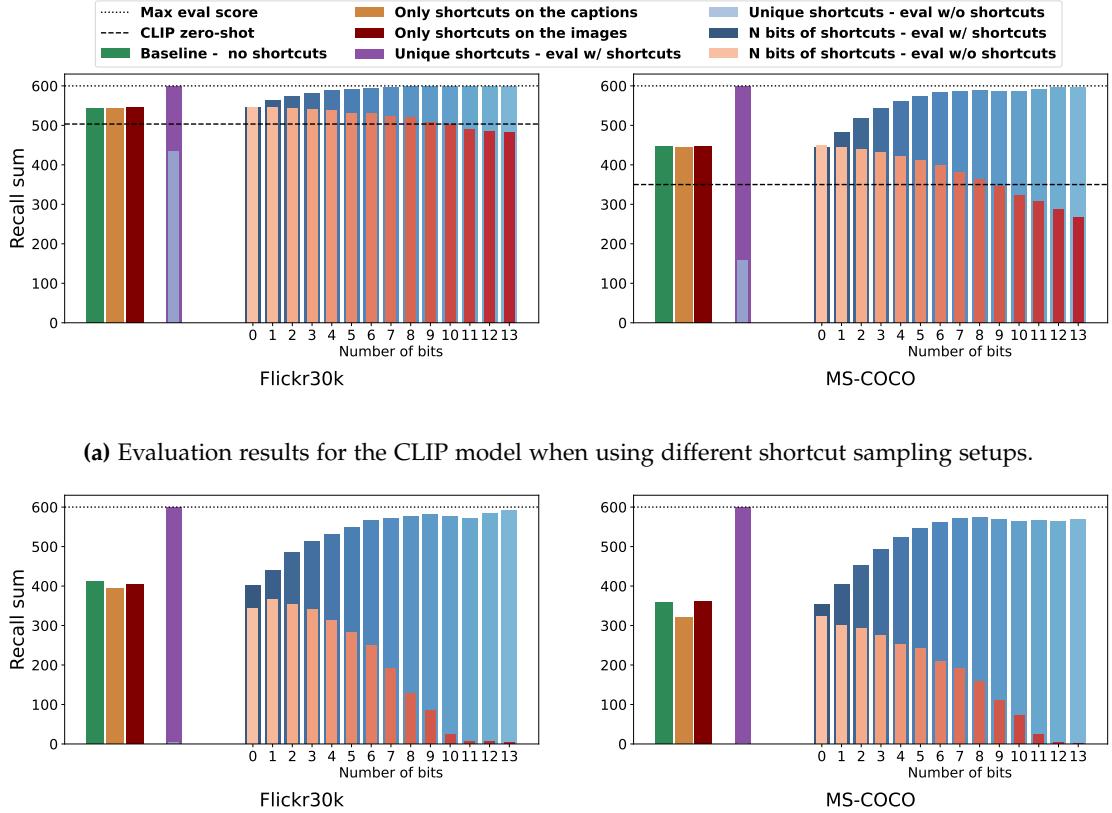


Figure 4.5: Effect of synthetic shortcuts on CLIP and VSE++ performance on ICR task. The dotted line represents the maximum achievable recall sum, while the dashed line for CLIP indicates its zero-shot evaluation performance. (Best viewed in color.)

We visualize the results in Figure 4.5. The dotted line (in Figure 4.5a and 4.5b) indicates the maximum evaluation score (i.e., 600). For CLIP, we also provide the zero-shot performance of the model, indicated by the dashed line in Figure 4.5a. When referring to specific results in Figure 4.5, we use the color of the corresponding bar and legend key in brackets in the text.

4.4.1 Findings

Based on Figure 4.5, we draw the following conclusions:

- I When training CLIP and VSE++ with only shortcuts on either the caption modality (in Figure 4.5, the corresponding bar/legend box is colored █) or on the image modality (█, in Figure 4.5), we do not observe a drop in evaluation scores for CLIP compared to the baseline model (█, in Figure 4.5a). For VSE++ we only observe a slight drop in evaluation score when training with shortcuts on the caption modality (again █, mainly for MS-COCO, in Figure 4.5b). There-

fore, we conclude that the synthetic shortcuts do not interfere with the original shared information S or other task-relevant information.

- II When training the models with *unique shortcuts*, we observe for both CLIP and VSE++ that when evaluating with shortcuts (■, in Figure 4.5), the models obtain a perfect evaluation score. When evaluating without shortcuts (■, in Figure 4.5) the evaluation score for VSE++ drops to zero and for CLIP below the zero-shot performance. We conclude that with unique shortcuts: (i) both CLIP and VSE++ fully rely on the shortcuts to solve the evaluation task, (ii) VSE++ has not learned any other shared or task-relevant information other than the shortcuts (since it is trained from scratch, only detecting the shortcuts is sufficient to minimize the contrastive loss), and (iii) fine-tuned CLIP has suppressed original features from the zero-shot model in favor of the shortcuts.
- III When training the models with N bits of *shortcuts*, we observe for both CLIP and VSE++ that the larger the number of bits we use during training and when evaluating without shortcuts (■, in Figure 4.5), the bigger the drop in evaluation performance. When we evaluate with shortcuts (■, in Figure 4.5), the evaluation performance improves as we use more bits compared to the baseline without shortcuts (■, in Figure 4.5). For VSE++, evaluating without shortcuts (■, in Figure 4.5b) results in a drop to zero when having a large number of bits. For CLIP, the evaluation performance drops below the zero-shot performance. If we train with 0 bits of shortcuts (i.e., the shortcut is a constant) we do not observe any drop or increase in evaluation scores for CLIP.

4.4.2 Upshot

Given the findings based on Figure 4.5 we conclude that a contrastive loss (i.e., InfoNCE) mainly learns easily detectable minimal features shared among pairs of images and captions. The learned features are sufficient to minimize the contrastive objective while suppressing the remaining shared and/or task-relevant information. If contrastive losses are sufficient to learn task-optimal representations for image-caption matching, these shortcuts should not adversely impact the evaluation performance. Moreover, if the contrastive loss would only learn features that are shared among the image and all captions (i.e, S), we should not observe a drop in performance to 0 for the VSE++ model when training with unique shortcuts, since there is still a lot of task-relevant information present in S . Especially in a training setup where a model is trained from scratch or fine-tuned on small datasets, the easy-to-detect features are likely not equivalent to all task-relevant information in the images and captions. Hence, we conclude that contrastive loss itself is not sufficient to learn task-optimal

representations of the images (and sufficient representations of captions) and that it only learns the minimal easy-to-detect features that are needed to minimize the contrastive objective.

4.5 REDUCING SHORTCUT LEARNING

In the earlier section, we have demonstrated that contrastive loss mainly relies on the minimal, easy-to-detect features shared among image-caption pairs while suppressing remaining task-relevant information. In this section, we describe two methods that help to reduce shortcut learning for contrastive learning on our SVL framework: latent target decoding (Bleeker et al., 2022) and implicit feature modification (Robinson et al., 2021).

4.5.1 Latent Target Decoding

Latent target decoding (LTD) (Bleeker et al., 2022) is a method to reduce predictive feature suppression (i.e., shortcut learning) for resource-constrained contrastive image-caption matching. The contrastive objective (i.e., InfoNCE) is combined with an additional reconstruction loss, which reconstructs the input caption from the latent representation of the caption $\mathbf{z}_{\mathcal{C}_j}^i$. We refer to Appendix 4.E.2 for the mathematical definition of LTD. Instead of reconstructing the tokens of the input caption in an auto-regressive manner (i.e., auto-encoding), the caption is reconstructed non-auto-regressively, by mapping the caption representation into the latent space of a Sentence-BERT (Reimers and Gurevych, 2019; Song et al., 2020) and minimizing the distance (i.e., reconstructing) between the reconstruction and the Sentence-BERT representation of the caption $\mathbf{x}_{\mathcal{C}_j}^i$. The assumption is that the *target* generated by the Sentence-BERT model contains all task-relevant information in the caption. Hence, by correctly mapping the latent caption representation $\mathbf{z}_{\mathcal{C}_j}^i$ into the latent space of Sentence-BERT, the caption encoder cannot suppress any task-relevant information or rely on shortcut solutions. LTD is implemented both as a dual-loss objective (i.e., the contrastive loss and LTD are added up) and as an optimization constraint while minimizing the InfoNCE loss, by implementing the loss as a Lagrange multiplier.

Experimental Setup. We use the LTD implementation and set-up similar to Bleeker et al. (2022). We train both CLIP and VSE++ with LTD, implemented as either dual loss or an optimization constraint. When implementing LTD as a constraint, we try $\eta \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ as bound values.

Similar to (Bleeker et al., 2022), when implementing LTD as a dual loss, we use $\beta = 1$

as balancing parameters. We train both with and without unique shortcuts. We do this to show (i) what the performance improvement is compared to using only InfoNCE, and (ii) to what degree LTD prevents full collapse to shortcut features. For each model and dataset, we take the training setup that results in the highest performance on the validation set.

4.5.2 *Implicit Feature Modification*

Implicit feature modification (IFM) (Robinson et al., 2021) is a method, originally introduced in the context of representation learning for images, that applies perturbations to logits used for guiding contrastive models. IFM perpetuates features that the encoders use during a training step to discriminate between positive and negative samples. By doing so, IFM alters the features that are currently used to solve the discrimination task, to avoid the InfoNCE loss to learn shortcut solutions. How much of the features are removed, is defined by a perturbation budget ϵ . IFM is implemented as a dual loss in combination with the InfoNCE loss. For the mathematical definition of IFM, we refer to Appendix 4.E.3.

Experimental Setup. We apply a similar experimental set-up for IFM as for LTD. We apply IFM both to CLIP and to VSE++, both with and without unique shortcuts. Similar to (Robinson et al., 2021), we try different perturbation budgets ϵ , we try $\epsilon \in \{0.05, 0.1, 0.2, 0.5, 1\}$. In line with the LTD setup, we take the training setup that results in the highest performance on the validation set.

4.5.3 *Method Comparison*

Both LTD and IFM aim to mitigate shortcut learning through different approaches. LTD aims to learn all task-relevant information by reconstructing the input captions. In contrast, IFM perturbs the discriminative features in the latent space of the encoder and does not rely on a reconstruction objective. Overall, both methods represent distinct strategies for improving the robustness and generalization capabilities of VL representation learning.

In the following section, we present experimental results with LTD and IFM, providing insight into their effectiveness in mitigating shortcut learning.

4.6 EXPERIMENTAL RESULTS

4.6.1 Does Latent Target Decoding Reduce Shortcut Learning?

In Table 4.1 we summarize the effect of LTD on reducing shortcut learning.

For CLIP, for both the Flickr30k and MS-COCO dataset, we do not observe an increase in recall scores when fine-tuning with $\mathcal{L}_{\text{InfoNCE+LTD}}$ compared to models that are only fine-tuned with $\mathcal{L}_{\text{InfoNCE}}$. LTD has originally been proposed for resource-constrained VL models. We argue that the additional features that LTD can extract are either already present in the pre-trained CLIP model, or not relevant for the evaluation task. However, when fine-tuning with $\mathcal{L}_{\text{InfoNCE+LTD}}$ and in the presence of shortcuts in the training data, degradation in recall scores is significantly lower than when fine-tuned only with the $\mathcal{L}_{\text{InfoNCE}}$. This shows that LTD can reduce the suppression of features in favor of the shortcut features when fine-tuning large-scale VL models.

Across the board, VSE++ models trained with the $\mathcal{L}_{\text{InfoNCE+LTD}}$ loss consistently outperform the $\mathcal{L}_{\text{InfoNCE}}$ loss, both for i2t and t2i retrieval and both when trained either with or without shortcuts, as indicated by higher recall@ k scores; this is consistent with the findings presented in (Bleeker et al., 2022)). For both the Flickr30k and MS-COCO dataset, when trained with the $\mathcal{L}_{\text{InfoNCE}}$ and with shortcuts present in the training data, the model performance collapses to around 0 in the absence of shortcuts (as we have seen in Section 4.4). However, when we train with shortcuts in the training data and with $\mathcal{L}_{\text{InfoNCE+LTD}}$, we observe, for both Flickr30k and MS-COCO, a significant gain in performance. The performance improvement is bigger for Flickr30k than for MS-COCO. In general, the recall scores are still significantly lower than training without shortcuts, however, the models do not solely rely on the shortcuts anymore to minimize the contrastive loss and are able during evaluation (in the absence of shortcuts) to still correctly match image-caption pairs with each other. The results in Table 4.1 show that LTD is able, in the presence of shortcuts in the training data, to guide (small-scale) VL models that are trained from scratch to not only learn the shortcut features that minimize the contrastive training objective but also represent other remaining task-relevant features in the data that are not extracted by $\mathcal{L}_{\text{InfoNCE}}$.

4.6.2 Does Implicit Feature Modification Reduce Shortcut Learning?

In Table 4.2 we summarize the effect of IFM on reducing shortcut solutions.

For CLIP, we observe that $\mathcal{L}_{\text{InfoNCE+IFM}}$, when training without shortcuts in the training data, only improves performance for the MS-COCO dataset for the t2i task. However, for both Flickr30k and MS-COCO we observe that, when training with unique shortcuts in the training data, fine-tuning with $\mathcal{L}_{\text{InfoNCE+IFM}}$ results in a significantly

Table 4.1: Mean and variance (over three training runs) recall@ k evaluation scores for the Flickr30k and MS-COCO datasets for image-to-text and text-to-image retrieval. We train with two loss functions: $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{InfoNCE+LTD}}$. We train either with (✓) or without (✗) shortcuts. For the model trained with $\mathcal{L}_{\text{InfoNCE+LTD}}$, we provide the hyper-parameters of the best-performing model. η indicates that the best-performing model uses LTD implemented as an optimization constraint with bound η . β indicates that the best-performing model uses LTD implemented as a dual-loss with $\beta = 1$.

Loss	S_{SynSC}	i2t			t2i			rsum		
		R@1	R@5	R@10	R@1	R@5	R@10			
Flickr30k										
CLIP										
$\mathcal{L}_{\text{InfoNCE}}$	✗	86.9 ± 0.1	97.4 ± 0.1	99.0 ± 0.0	72.4 ± 0.1	92.1 ± 0.0	95.8 ± 0.0	543.5 ± 1.1		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \beta = 1$	✗	86.5 ± 0.6	97.1 ± 0.0 ↓	98.5 ± 0.0 ↓	72.4 ± 0.0 —	92.3 ± 0.0 ↓	95.9 ± 0.0 ↓	542.8 ± 0.8 —		
$\mathcal{L}_{\text{InfoNCE}}$	✓	57.2 ± 8.3	84.0 ± 4.8	91.0 ± 1.9	44.9 ± 4.5	74.9 ± 6.0	84.2 ± 2.5	436.2 ± 145.0		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \beta = 1$	✓	64.0 ± 1.3 ↑	87.8 ± 0.9 ↑	93.2 ± 0.8 ↑	50.7 ± 0.6 ↑	79.8 ± 0.7 ↑	88.1 ± 0.5 ↑	463.6 ± 17.3 ↑		
VSE++										
$\mathcal{L}_{\text{InfoNCE}}$	✗	52.6 ± 1.1	79.8 ± 0.1	87.8 ± 0.1	39.5 ± 0.3	69.8 ± 0.0	79.4 ± 0.1	409.0 ± 4.0		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \eta = 0.2$	✗	54.1 ± 0.1 ↑	81.1 ± 0.8 ↑	88.6 ± 0.1 ↑	42.5 ± 0.0 ↑	71.9 ± 0.1 ↑	81.3 ± 0.0 ↑	419.6 ± 0.1 ↑		
$\mathcal{L}_{\text{InfoNCE}}$	✓	0.1 ± 0.0	0.6 ± 0.1	1.1 ± 0.1	0.1 ± 0.0	0.5 ± 0.0	1.0 ± 0.0	3.4 ± 0.6		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \eta = 0.05$	✓	24.7 ± 0.5 ↑	51.8 ± 0.7 ↑	65.6 ± 1.4 ↑	20.7 ± 1.0 ↑	49.2 ± 0.6 ↑	62.6 ± 1.2 ↑	274.6 ± 4.6 ↑		
MS-COCO										
CLIP										
$\mathcal{L}_{\text{InfoNCE}}$	✗	63.8 ± 0.3	86.1 ± 0.2	92.3 ± 0.0	46.3 ± 0.3	74.8 ± 0.1	84.1 ± 0.2	447.5 ± 0.5		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \beta = 1$	✗	63.8 ± 0.0 —	86.1 ± 0.0 —	92.3 ± 0.0 —	46.3 ± 0.0 —	74.7 ± 0.0 —	84.1 ± 0.0 —	447.4 ± 0.0 —		
$\mathcal{L}_{\text{InfoNCE}}$	✓	13.6 ± 0.9	31.5 ± 2.4	42.2 ± 3.7	7.3 ± 0.6	22.1 ± 1.0	32.7 ± 1.7	149.4 ± 32.7		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \beta = 1$	✓	18.9 ± 0.1 ↑	41.8 ± 0.1 ↑	54.1 ± 0.1 ↑	16.5 ± 0.0 ↑	39.4 ± 0.0 ↑	52.6 ± 0.1 ↑	223.4 ± 0.2 ↑		
VSE++										
$\mathcal{L}_{\text{InfoNCE}}$	✗	42.2 ± 0.1	72.7 ± 0.1	83.2 ± 0.1	30.9 ± 0.0	61.2 ± 0.1	73.5 ± 0.1	363.8 ± 2.3		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \eta = 0.1$	✗	43.6 ± 0.1 ↑	73.5 ± 0.0 ↑	83.7 ± 0.0 ↑	32.4 ± 0.1 ↑	62.5 ± 0.0 ↑	74.7 ± 0.0 ↑	370.5 ± 0.1 ↑		
$\mathcal{L}_{\text{InfoNCE}}$	✓	0.0 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.0 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.7 ± 0.0		
$\mathcal{L}_{\text{InfoNCE+LTD}}, \eta = 0.01$	✓	3.9 ± 0.0 ↑	13.7 ± 0.6 ↑	21.6 ± 0.9 ↑	3.1 ± 0.2 ↑	11.0 ± 1.6 ↑	18.1 ± 3.0 ↑	71.3 ± 3.6 ↑		

lower performance drop in recall score than when fine-tuning with the $\mathcal{L}_{\text{InfoNCE}}$. Similar to LTD, the recall@ k scores are still lower than when trained without shortcuts in the training data. We conclude that IFM is sufficient to reduce the suppression of features in favor of the shortcut features when fine-tuning a large-scale VL model, as indicated by higher recall@ k scores when evaluating without shortcuts.

For VSE++, both for the Flickr30k and MS-COCO dataset, we do not observe that $\mathcal{L}_{\text{InfoNCE+IFM}}$ outperforms the $\mathcal{L}_{\text{InfoNCE}}$, both with and without shortcuts present in

Table 4.2: Mean and variance (over three training runs) recall@ k evaluation scores for the Flickr30k and MS-COCO datasets for image-to-text and text-to-image retrieval. We train with two loss functions: $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{InfoNCE+IFM}}$. We train either with (✓) or without (✗) shortcuts. For the model trained with $\mathcal{L}_{\text{InfoNCE+IFM}}$, we provide the hyper-parameters of the best-performing model.

Loss	S_{SynSC}	$i2t$			$t2i$			rsum		
		R@1	R@5	R@10	R@1	R@5	R@10			
Flickr30k										
CLIP										
$\mathcal{L}_{\text{InfoNCE}}$	✗	86.9 ± 0.1	97.4 ± 0.0	98.8 ± 0.0	72.8 ± 0.2	92.1 ± 0.0	95.6 ± 0.0	543.5 ± 1.3		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✗	$87.4 \pm 0.1 \uparrow$	$97.4 \pm 0.2 \downarrow$	$99.1 \pm 0.0 \downarrow$	$73.2 \pm 0.0 \downarrow$	$92.2 \pm 0.0 \downarrow$	$95.6 \pm 0.0 \downarrow$	$544.9 \pm 0.2 \downarrow$		
$\mathcal{L}_{\text{InfoNCE}}$	✓	57.9 ± 0.3	84.6 ± 0.8	91.3 ± 0.0	43.9 ± 2.2	74.6 ± 0.8	84.4 ± 0.4	436.7 ± 18.8		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.1$	✓	$73.8 \pm 0.8 \uparrow$	$91.5 \pm 0.5 \uparrow$	$95.6 \pm 0.0 \uparrow$	$58.9 \pm 0.1 \uparrow$	$84.4 \pm 0.1 \uparrow$	$91.1 \pm 0.2 \uparrow$	$495.2 \pm 5.7 \uparrow$		
VSE++										
$\mathcal{L}_{\text{InfoNCE}}$	✗	52.9 ± 0.2	80.5 ± 0.1	87.6 ± 0.4	40.5 ± 0.1	68.8 ± 0.4	78.9 ± 0.3	409.3 ± 2.6		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✗	$52.4 \pm 0.2 \downarrow$	$76.9 \pm 0.1 \downarrow$	$85.3 \pm 0.0 \downarrow$	$39.1 \pm 0.0 \downarrow$	$68.8 \pm 0.1 \downarrow$	$78.2 \pm 0.1 \downarrow$	$400.7 \pm 0.0 \downarrow$		
$\mathcal{L}_{\text{InfoNCE}}$	✓	0.1 ± 0.0	0.4 ± 0.0	0.8 ± 0.0	0.1 ± 0.0	0.4 ± 0.0	1.0 ± 0.0	2.9 ± 0.0		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✓	$0.0 \pm 0.0 \downarrow$	$0.6 \pm 0.1 \downarrow$	$0.9 \pm 0.2 \downarrow$	$0.1 \pm 0.0 \downarrow$	$0.5 \pm 0.0 \downarrow$	$1.0 \pm 0.0 \downarrow$	$3.2 \pm 0.8 \downarrow$		
MS-COCO										
CLIP										
$\mathcal{L}_{\text{InfoNCE}}$	✗	63.5 ± 0.1	86.0 ± 0.3	92.2 ± 0.0	46.3 ± 0.0	74.7 ± 0.0	84.2 ± 0.0	446.9 ± 0.9		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✗	$63.0 \pm 0.1 \downarrow$	$86.6 \pm 0.1 \downarrow$	$92.6 \pm 0.2 \downarrow$	$47.2 \pm 0.0 \uparrow$	$75.6 \pm 0.0 \uparrow$	$84.5 \pm 0.0 \uparrow$	$449.5 \pm 1.7 \uparrow$		
$\mathcal{L}_{\text{InfoNCE}}$	✓	13.9 ± 0.0	32.7 ± 0.1	43.8 ± 0.0	8.8 ± 0.0	24.7 ± 0.2	35.5 ± 0.5	159.4 ± 3.4		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✓	$23.4 \pm 1.5 \uparrow$	$46.5 \pm 2.7 \uparrow$	$58.2 \pm 2.5 \uparrow$	$17.1 \pm 0.3 \uparrow$	$38.9 \pm 0.9 \uparrow$	$51.3 \pm 1.0 \uparrow$	$235.5 \pm 43.8 \uparrow$		
VSE++										
$\mathcal{L}_{\text{InfoNCE}}$	✗	41.7 ± 0.3	72.5 ± 0.1	83.1 ± 0.1	31.3 ± 0.0	61.1 ± 0.0	73.6 ± 0.0	363.4 ± 0.4		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✗	$40.2 \pm 0.0 \downarrow$	$70.8 \pm 0.1 \downarrow$	$81.6 \pm 0.1 \downarrow$	$30.8 \pm 0.0 \downarrow$	$61.5 \pm 0.0 \uparrow$	$74.3 \pm 0.0 \uparrow$	$359.3 \pm 1.1 \downarrow$		
$\mathcal{L}_{\text{InfoNCE}}$	✓	0.0 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.0 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.6 ± 0.0		
$\mathcal{L}_{\text{InfoNCE+IFM}}, \epsilon = 0.05$	✓	$0.0 \pm 0.0 \downarrow$	$0.1 \pm 0.0 \downarrow$	$0.2 \pm 0.0 \downarrow$	$0.0 \pm 0.0 \downarrow$	$0.1 \pm 0.0 \downarrow$	$0.2 \pm 0.0 \downarrow$	$0.7 \pm 0.0 \downarrow$		

the training data. We even observe that $\mathcal{L}_{\text{InfoNCE+IFM}}$, when training without shortcuts, results in a decrease in performance across all recall@ k metrics. When training with $\mathcal{L}_{\text{InfoNCE+IFM}}$ and with unique shortcuts in the training data, the evaluation performance still collapses to around 0. The results in Table 4.2 show that IFM is not sufficient to prevent models trained from scratch from fully collapsing to the artificial shortcut solutions we introduce in this chapter (as opposed to LTD).

4.6.3 Upshot

In this section, we have evaluated two methods for reducing shortcut learning on our SVL framework: LTD and IFM. LTD proves effective in reducing shortcut learning for both CLIP and VSE++. IFM demonstrates its efficacy solely during the fine-tuning of CLIP. These findings indicate that our SVL framework is a challenging and interesting framework to study and evaluate shortcut learning for contrastive VL models. Moreover, our results show that shortcut learning is only partially addressed by the evaluated methods since the evaluation results are not on par with the results on data lacking synthetic shortcuts.

4.7 RELATED WORK

We discuss related work on multi-view representation learning, vision-language learning, and shortcut learning.

4.7.1 Multi-View Representation Learning

To learn the underlying semantics of the training data, a subgroup of representation learning methods involves training neural encoders that maximize the agreement between representations of the similar *views* (Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020a; Radford et al., 2021; Bardes et al., 2022). In general, for uni-modal representation learning, data augmentations are used to generate different views of the same data point. One of the core assumptions in multi-view representation learning is that each view shares the same *task-relevant information* (Sridharan and Kakade, 2008; Zhao et al., 2017; Federici et al., 2020; Tian et al., 2020a; Shwartz-Ziv and LeCun, 2023). However, the optimal view for contrastive self-supervised learning (SSL) (i.e., which information is shared among views/which data augmentation is used) is task-dependent (Tian et al., 2020b; Xiao et al., 2021). Therefore, maximizing the mutual information (MI) between representations of views (i.e., shared information) does not necessarily result in representations that generalize better to down-stream evaluation tasks, since the representations may contain too much additional noise that is irrelevant for the downstream task (Tian et al., 2020b; Tschannen et al., 2020). An open problem in multi-view SSL is to learn representations that contain all task-relevant information from views where each view contains distinct, task-relevant information (Shwartz-Ziv and LeCun, 2023), this is especially a problem in the multi-modal learning domain (Zong et al., 2023).

Chen et al. (2021) investigate multi-view representation learning for images using

contrastive losses. They demonstrate that when multiple competing features exist that redundantly predict the match between two views, contrastive models tend to focus on learning the easy-to-represent features while suppressing other task-relevant information. This results in contrastive losses mainly capturing the easy features, even if all task-relevant information is shared between the two views, suppressing the remaining relevant information.

Several optimization objectives have been introduced to either maximize the lower bound on the MI between views and their latent representations (Oord et al., 2018; Bachman et al., 2019; Hjelm et al., 2019; Tian et al., 2020a) or minimize the MI between representations of views while keeping the task-relevant information (Federici et al., 2020; Lee et al., 2021). To learn more task-relevant information that either might not be shared between views or that is compressed by a contrastive loss, several works proposed additional reconstruction objectives to maximize the MI between the latent representation and input data (Tsai et al., 2021; Wang et al., 2022a; Li et al., 2023b; Bleeker et al., 2022). Liang et al. (2023) introduce a multimodal contrastive objective that factorizes the representations into shared and unique information, while also removing task-irrelevant information by minimizing the upper bound on MI between similar views.

4.7.2 Vision-language Representation Learning

The goal of VL representation learning is to combine information from the visual and textual modalities into a joint representation or learn coordinated representations (Baltrusaitis et al., 2019; Guo et al., 2019b). The representation learning approaches can be separated into several groups.

Contrastive methods represent one prominent category of VL representation methods. The approaches in this group are typically dual encoders. Early methods in this category are trained from scratch; for instance, Frome et al. (2013) proposed a VL representation learning model that features a skip-gram language model and a visual object categorization component trained with hinge rank loss. Another subgroup of methods uses a *dual-encoder* with a hinge-based triplet loss (Kiros et al., 2014; Li et al., 2019a; Lee et al., 2018). (Kiros et al., 2014) use the loss for training a CNN-RNN dual encoder. Li et al. (2019a) leverage bottom-up attention and graph convolutional networks (Kipf and Welling, 2017) to learn the relationship between image regions. Lee et al. (2018) add stacked cross-attention to use both image regions and words as context.

Recently, contrastive approaches involve transformer-based dual-encoders trained with more data than the training data from the evaluation set(s). ALBEF (Li et al., 2021a) propose to contrastively align unimodal representations before fusion, while

X-VLM (Zeng et al., 2022) employs an additional cross-modal encoder to learn fine-grained VL representations. Florence (Yuan et al., 2021) leverages various adaptation models for learning fine-grained object-level representations. CLIP (Radford et al., 2021), a scaled-up dual-encoder, is pre-trained on the task of predicting which caption goes with which image. ALIGN (Jia et al., 2021) uses a simple dual-encoder trained on over a billion image alt-text pairs. FILIP (Yao et al., 2022) is a transformer-based bi-encoder that features late multimodal interaction meant to capture fine-grained representations. SLIP (Mu et al., 2022) combines language supervision and image self-supervision to learn visual representations without labels. DeCLIP (Li et al., 2022c) proposes to improve the efficiency of CLIP pretraining using intra-modality self-supervision, cross-modal multi-view supervision, and nearest neighbor supervision.

Another line of work includes learning VL representations using models that are inspired by BERT (Devlin et al., 2019). ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) expand upon BERT by introducing a two-stream architecture, where two transformers are applied to images and text independently, which is fused by a third transformer in a later stage. B2T2 (Alberti et al., 2019), VisualBERT (Li et al., 2019b), Unicoder-VL (Li et al., 2020a), VL-BERT (Su et al., 2020), and UNITER (Chen et al., 2020b) propose a single-stream architecture, where a single transformer is applied to both images and text. Oscar (Li et al., 2020d) uses caption object tags as anchor points that are fed to the transformer alongside region features. BEIT-3 (Wang et al., 2022b) adapt multiway transformers trained using cross-entropy loss (Bao et al., 2022).

Another category of methods for learning VL representations are generative methods, that imply learning VL representation by generating new instances of one modality conditioned on the other modality. For instance, BLIP (Li et al., 2022b) bootstraps captions by generating synthetic captions and filtering out the noisy ones; BLIP-2 (Li et al., 2023a) bootstraps VL representation learning and, subsequently, vision-to-language generative learning. On the other hand, Tschannen et al. (2023) propose to pretrain a encoder-decoder architecture via the image captioning task.

Shortcut Learning. Geirhos et al. (2020) define shortcuts in deep neural networks as “decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios.” In the context of deep learning, a shortcut solution can also be seen as a discrepancy between the features that a model has learned during training and the intended features that a model should learn to perform well during evaluation. For example, shortcuts might be features that minimize the training objective but are much easier to detect than the intended features that are relevant to the evaluation task. Shortcut learning can be caused by biases in the dataset or inductive biases in either the network architecture or training objective.

Hermann and Lampinen (2020) design a dataset with multiple predictive features, where each feature can be used as a label for an image classification task. The authors show that in the presence of multiple features that each redundantly predicts the target label, the deep neural model chooses to represent only one of the predictive features that are the easiest to detect, i.e., the model favors features that are easy to detect over features that are harder to discriminate. Next to that, they show that features that are not needed for a classification task, are in general suppressed by the model instead of captured in the learned latent representations.

Robinson et al. (2021) show that contrastive losses can have multiple local minima, where different local minima can be achieved by suppressing features from the input data (i.e., the model learns a shortcut by not learning all task-relevant features). To mitigate the shortcut learning problem, Robinson et al. (2021) propose implicit feature modification, a method that perpetuates the features of positive and negative samples during training to encourage the model to capture different features than the model currently relies on.

Scimeca et al. (2022) design an experimental set-up with multiple shortcut cues in the training data, where each shortcut is equally valid w.r.t. predicting the correct target label. The goal of the experimental setup is to investigate which cues are preferred to others when learning a classification task.

Latent target decoding (LTD) is a method to reduce predictive feature suppression (i.e., shortcuts) for resource-constrained contrastive ICR by reconstructing the input caption in a non-auto-regressive manner. Bleeker et al. (2022) argue that most of the task-relevant information for the ICR task is captured by the text modality. Hence, the focus is on the reconstruction of the text modality instead of the image modality. Bleeker et al. (2022) add a decoder to the learning algorithm, to reconstruct the input caption. Instead of reconstructing the input tokens, the input caption is reconstructed in a non-autoregressive manner in the latent space of a Sentence-BERT (Reimers and Gurevych, 2019; Song et al., 2020) model. LTD can be implemented as an optimization constraint and as a dual-loss. Li et al. (2023b) show that contrastive losses are prone to feature suppression. They introduce predictive contrastive learning (PCL), which combines contrastive learning with a decoder to reconstruct the input data from the latent representations to prevent shortcut learning.

Adnan et al. (2022) measure the MI between the latent representation and the input as a domain agnostic metric to find where (and when) in training the neural network relies on shortcuts in the input data. Their main finding is that, in the presence of a shortcut, the MI between the input data and the latent representation of the data is lower than without a shortcut in the input data. Hence, the latent representation captures less information of the input data in the presence of the shortcut and mainly relies on the shortcut to predict the target.

4.7.3 Our Focus

In this chapter, we focus on the problem of shortcut learning for VL in the context of multi-view VL representation learning with multiple captions per image. In contrast with previous (uni-modal) work on multi-view learning, we consider different captions matching to the same image as different *views*. We examine the problem by introducing a framework of synthetic shortcuts designed for VL representation learning, which allows us to investigate the problem in a controlled way. For our experiments, we select two prevalent VL models that are solely optimized with the InfoNCE loss: CLIP, a large-scale pre-trained model, and VSE++, a model trained from scratch. We select models that are solely optimized with a contrastive loss, to prevent measuring the effect of other optimization objectives on the shortcut learning problem.

4.8 CONCLUSION

In this chapter, we focus on the shortcut learning problem of contrastive learning in the context of vision-language (VL) representation learning with multiple captions per image. We have proposed synthetic shortcuts for vision-language (SVL): a training and evaluation framework to examine the problem of shortcut learning in a controlled way. The key component of this framework is synthetic shortcuts that we add to image-text data. Synthetic shortcuts represent additional, easily identifiable information that is shared between images and captions. We fine-tune CLIP and train a VSE++ model from scratch using our training framework to evaluate how prone contrastive VL models are to shortcut learning. Next, we have evaluated how shortcut learning can be partially mitigated using latent target decoding and implicit feature modification.

Main Findings. We have conducted experiments on two distinct VL models, CLIP and VSE++, and have evaluated the performance on Flickr30k and MS-COCO. We have found that when training with unique shortcuts, CLIP suppresses pre-trained features in favor of the shortcuts. VSE++ only learns to represent the shortcuts, when using unique shortcuts, showing that none of the remaining task-relevant (both shared and unique) information is captured by the encoders when training a model from scratch. When using n bits of *shortcuts*, we have shown that the more bits we use, the more the contrastive VL models rely on the synthetic shortcuts. Our results demonstrate that contrastive VL methods tend to depend on easy-to-learn discriminatory features shared among images and all matching captions while suppressing the remaining task-relevant information. Next, we have evaluated two methods for reducing shortcut learning on our framework of synthetic shortcuts for image-caption datasets. Both

methods partially mitigate shortcut learning when training and evaluating with our shortcut learning framework. These findings show that our framework is a challenging framework to study and evaluate shortcut learning for contrastive VL and underline the complexity of our framework in studying and evaluating shortcut learning within the context of contrastive VL representation learning.

Implications. The implications of our findings are twofold. First, we examine the limitations of contrastive optimization objectives for VL representation learning, demonstrating that they predominantly capture features that are easily discriminable but may not necessarily constitute task-optimal representations. Second, our work contributes a novel framework for investigating shortcut learning problem in the context of VL representation learning with multiple captions per image, providing insights into the extent to which models rely on shortcuts when they are available and how existing shortcut reduction methods are capable of reducing shortcut learning when training with our framework.

Limitations. Some of the limitations of our work are related to the fact that we focused on two specific models, one optimization objective (InfoNCE), and two datasets, and the generalizability of our findings to other VL models, optimization objectives, and datasets warrants further exploration. Additionally, the synthetic shortcuts introduced in this chapter are not dependent on image-caption pairs. Our training and evaluation setup shows that, in the presence of shortcuts in the training data, contrastive VL models mainly rely on the easy-to-detect shortcut features, which indicates that the InfoNCE loss cannot learn tasks-optimal representations for VL tasks when multiple captions are used for training. However, it remains unclear to what degree the unique information of the captions is captured by the contrastive loss VL models.

Future Work. We suggest working on the development of optimization objectives that specifically address the shortcut learning problem for VL training with multiple captions per image. Moreover, we suggest extending our synthetic shortcuts for image-caption datasets to a framework with unique shortcut information per caption. By having unique shortcut information per caption, it becomes possible to measure how much of the shared/caption-specific shortcut information is captured by the encoder models. Another direction for future research includes investigating alternative training strategies or loss functions to further mitigate shortcut learning problems. Another promising direction for future work includes the improvement of existing methods or the exploration of novel techniques that address the limitations of existing shortcut reduction methods, potentially through the combination of multiple approaches. Finally, extending the SVL framework to better capture nuances and complexities of natural data is another important and promising direction. This would allow a more comprehensive exploration of shortcut learning and the understanding of the implications in

real-world scenarios and datasets.

Answer to RQ3. Hence, our findings indicate that in vision-language representation learning with multiple captions per image, the presence of shortcuts hinders the learning of task-optimal representations. We assume that this happens because contrastive learning approaches prioritize easily detectable features shared between the image and all captions, neglecting unique information specific to individual captions. This dependence on shortcuts prevents models from capturing the full spectrum of relevant information within the image and its captions, resulting in suboptimal representations.

4.9 BROADER IMPACT

This chapter motivates and introduces a framework for investigating the problem of shortcut learning for contrastive VL representation learning with multiple captions per image in a controlled way. It also examines how two shortcut learning reduction methods perform on the proposed framework. Overall, the framework provides a tool for analyzing and understanding the problem of shortcut learning in the context of contrastive VL representation learning; it can be used in various settings that require deeper insight into the quality of learned VL representations.

We should be aware that the reliance on shortcuts in VLMs poses ethical concerns with potential real-world implications. Models that learn shortcuts may overlook nuanced details in images and text, leading to biased or inaccurate outcomes. Furthermore, transparency and explainability of VLMs are crucial considerations. Models that rely on shortcuts may make decisions based on features that are not easily interpretable or explainable to users. This lack of transparency can diminish trust in AI systems.

CHAPTER APPENDIX

4.A NOTATION

Symbol	Description
$\mathcal{L}_{\text{InfoNCE}}$	InfoNCE loss
$\mathcal{L}_{\text{InfoNCE+LTD}}$	Loss that combines InfoNCE and latent target decoding (LTD)
$\mathcal{L}_{\text{InfoNCE+IFM}}$	Loss that combines InfoNCE and implicit feature modification (IFM)
$\mathcal{L}_{\text{recon}}$	Reconstruction loss
\mathcal{D}	Dataset \mathcal{D} that comprises N image-caption tuples: $\mathcal{D} = \left\{ \left(\mathbf{x}_{\mathcal{I}}^i, \{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k \right) \right\}_{i=1}^N$; i -th image-caption tuple in the dataset \mathcal{D} consist out of an image $\mathbf{x}_{\mathcal{I}}^i$ and k associated captions $\{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k$
\mathcal{B}	Batch of image-caption pairs
$\mathbf{z}_{\mathcal{I}}$	Latent representation of image $\mathbf{x}_{\mathcal{I}}$
$\mathbf{z}_{\mathcal{C}}$	Latent representation of caption $\mathbf{x}_{\mathcal{C}}$
$\mathbf{z}_{\mathcal{C} \rightarrow \mathcal{I}}^{\text{SUF}}$	Latent representation of the caption $\mathbf{x}_{\mathcal{C}}$ that is sufficient for the image $\mathbf{x}_{\mathcal{I}}$
$\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{\text{SUF}}$	Latent representation of the image $\mathbf{x}_{\mathcal{I}}$ sufficient for the caption $\mathbf{x}_{\mathcal{C}}$
$\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{\text{MIN}}$	Latent representation of the image $\mathbf{x}_{\mathcal{I}}$ that is minimal sufficient for the caption $\mathbf{x}_{\mathcal{C}}$
$\mathbf{z}_{\mathcal{I} \rightarrow K}^{\text{OPT}}$	Latent representation of the image $\mathbf{x}_{\mathcal{I}}$ that is optimal for the set of captions K given the task T
$f_{\theta}(\cdot)$	Image encoder parametrised by θ ; takes image $\mathbf{x}_{\mathcal{I}}$ as input and returns its latent representation $\mathbf{z}_{\mathcal{I}}$: $\mathbf{z}_{\mathcal{I}} := f_{\theta}(\mathbf{x}_{\mathcal{I}})$
$g_{\phi}(\cdot)$	Caption encoder parametrised by ϕ ; takes caption $\mathbf{x}_{\mathcal{C}}$ as input and returns its latent representation $\mathbf{z}_{\mathcal{C}}$: $\mathbf{z}_{\mathcal{C}} := g_{\phi}(\mathbf{x}_{\mathcal{C}})$
τ	Temperature parameter of $\mathcal{L}_{\text{InfoNCE}}$
ϵ	Perturbation budget for \mathcal{L}_{IFM}
η	Reconstruction bound for \mathcal{L}_{LTD}

Table 4.A.1: Overview of notation used in the chapter.

4.B PROBLEM DEFINITION AND ASSUMPTIONS

In this chapter, we solely focus on contrastive VL representation learning. We work in a setting where we investigate the problem by fine-tuning a large pre-trained foundation model (CLIP, Radford et al., 2021) and training a resource-constrained image-text method from scratch (VSE++, Faghri et al., 2018). We train and evaluate using two benchmark datasets where multiple captions per image are available: Flickr30k (Young et al., 2014) and MS-COCO Captions (Lin et al., 2014). Both datasets come with 5 captions per image. We work in a dual-encoder setup, i.e., we have a separate image and caption encoder, which do not share parameters.

4.B.1 Evaluation Task

The image-caption retrieval (ICR) evaluation task consists of two sub-tasks: image-to-text (i2t) and text-to-image (t2i) retrieval. In ICR, either an image or a caption is used as a query and the goal is to rank a set of candidates in the other modality. In this chapter, we follow the standard ICR evaluation procedure (see, e.g., Faghri et al., 2018; Lee et al., 2018; Li et al., 2019a). The evaluation metric for the ICR task is Recall@ k , with $k = \{1, 5, 10\}$. For t2i retrieval, there is one matching/positive image per query caption (when using the Flickr30k or MS-COCO or dataset). Hence, the Recall@ k metric represents how often the correct image is present in the top- k of the ranking. For i2t retrieval, however, there are 5 matching captions per image. Therefore, only the highest-ranked correct caption is taken into account when measuring the Recall@ k (i.e., in the highest-ranked caption present in the top k). Standard practice to select the best model checkpoint during training is to use the *recall sum* (rsum) as a validation metric. The recall sum is the sum of recall at 1, 5, and 10, for both i2t and t2i. Therefore, the maximum value of the recall sum is 600.

4.B.2 Assumptions

Throughout this chapter, we rely on several assumptions about the problem definition. Our assumptions are defined at the level of an image-text tuple. Following Section 4.2, we formalize the assumptions on the case where one image is associated with two captions: $(\mathbf{x}_I, \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\})$.

Assumption 1. Each caption in the tuple contains information that is distinct from the other captions in the tuple and all captions and image in the tuple contain shared and unique information:

$$\begin{aligned} I(\mathbf{x}_I; \mathbf{x}_{C_A}; \mathbf{x}_{C_B}) &> 0 \\ I(\mathbf{x}_I; \mathbf{x}_{C_A} | \mathbf{x}_{C_B}) &> 0, \quad I(\mathbf{x}_I; \mathbf{x}_{C_B} | \mathbf{x}_{C_A}) > 0 \text{ and } I(\mathbf{x}_{C_A}; \mathbf{x}_{C_B} | \mathbf{x}_I) > 0 \\ H(\mathbf{x}_I | \mathbf{x}_{C_A}, \mathbf{x}_{C_B}) &> 0, \quad H(\mathbf{x}_{C_A} | \mathbf{x}_I, \mathbf{x}_{C_B}) > 0 \text{ and } H(\mathbf{x}_{C_B} | \mathbf{x}_I, \mathbf{x}_{C_A}) > 0. \end{aligned}$$

Assumption 2. Task-relevant information R is the combination of all the information shared between an image and each caption in the tuple:

$$R = I(\mathbf{x}_I; \mathbf{x}_{C_A} | \mathbf{x}_{C_B}) + I(\mathbf{x}_I; \mathbf{x}_{C_B} | \mathbf{x}_{C_A}) + I(\mathbf{x}_I; \mathbf{x}_{C_A}; \mathbf{x}_{C_B}).$$

4.C ANALYSIS OF CONTRASTIVE LEARNING FOR MULTIPLE CAPTIONS PER IMAGE

Theorem 1 (Suboptimality of contrastive learning with multiple captions per image). Given an image \mathbf{x}_I , a set of matching captions $\mathcal{C} = \{\mathbf{x}_{C_A}, \mathbf{x}_{C_B}\}$, and a contrastive learning loss function $\mathcal{L}_{\text{InfoNCE}}$ that optimizes for task T , image representations learned during contrastive learning will be minimal sufficient and will never be task-optimal image representations. More formally, assume that:

- (H1) $\forall i, j \in \{A, B\}$ such that $i \neq j$, $I(\mathbf{z}_{I \rightarrow C_i}^{\text{MIN}}; \mathbf{x}_{C_i}) = I(\mathbf{x}_I; \mathbf{x}_{C_i} | \mathbf{x}_{C_j}) + I(\mathbf{x}_I; \mathbf{x}_{C_i}; \mathbf{x}_{C_j})$.
- (H2) $\exists i, j \in \{A, B\}$ with $i \neq j$ such that $I(\mathbf{x}_I; \mathbf{x}_{C_i} | \mathbf{x}_{C_j}) > 0$.

Then the following holds:

$$(T_2) \quad \exists i \in \{A, B\} \text{ such that } I(\mathbf{z}_{I \rightarrow \mathcal{C}}^{\text{OPT}}; \mathbf{x}_{C_A}, \mathbf{x}_{C_B}) > I(\mathbf{z}_{I \rightarrow C_i}^{\text{MIN}}; \mathbf{x}_{C_i}).$$

Proof. Following Eq. 4.1 we define a task-optimal representation of an image \mathbf{x}_I w.r.t. all matching captions in \mathcal{C} as:

$$I(\mathbf{z}_{I \rightarrow \mathcal{C}}^{\text{OPT}}; \mathbf{x}_{C_A}, \mathbf{x}_{C_B}) = \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_A} | \mathbf{x}_{C_B})}_{C_A} + \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_B} | \mathbf{x}_{C_A})}_{C_B} + \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_A}; \mathbf{x}_{C_B})}_{S}.$$

Furthermore, following Definition 5, we define minimal sufficient representations of image \mathbf{x}_I w.r.t. each matching caption in \mathcal{C} as a combination of caption-specific and shared information:

$$I(\mathbf{z}_{I \rightarrow C_A}^{\text{MIN}}; \mathbf{x}_{C_A}) = \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_A} | \mathbf{x}_{C_B})}_{C_A} + \underbrace{I(\mathbf{x}_I; \mathbf{x}_{C_A}; \mathbf{x}_{C_B})}_{S}$$

$$I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}_B}^{MIN}; \mathbf{x}_{\mathcal{C}_B}) = \underbrace{I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_B} \mid \mathbf{x}_{\mathcal{C}_A})}_{C_B} + \underbrace{I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A}; \mathbf{x}_{\mathcal{C}_B})}_{S}.$$

Following assumption H_2 , for at least one caption $\mathbf{x}_C \in \mathcal{C}$ associated with the image $\mathbf{x}_{\mathcal{I}}$, caption-specific information is positive. Therefore, we consider two cases:

- If caption-specific information of $\mathbf{x}_{\mathcal{C}_A}$ is positive, that is, if $I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A} \mid \mathbf{x}_{\mathcal{C}_B}) > 0$:

$$\begin{aligned} \underbrace{I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A} \mid \mathbf{x}_{\mathcal{C}_B}) + I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_B} \mid \mathbf{x}_{\mathcal{C}_A}) + I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A}; \mathbf{x}_{\mathcal{C}_B})}_{(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{OPT}; \mathbf{x}_{\mathcal{C}_A} \mathbf{x}_{\mathcal{C}_B})} &> \underbrace{I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_B} \mid \mathbf{x}_{\mathcal{C}_A}) + I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A}; \mathbf{x}_{\mathcal{C}_B})}_{I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}_B}^{MIN}; \mathbf{x}_{\mathcal{C}_B})} \Rightarrow \\ &\Rightarrow I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{OPT}; \mathbf{x}_{\mathcal{C}_A} \mathbf{x}_{\mathcal{C}_B}) > I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}_B}^{MIN}; \mathbf{x}_{\mathcal{C}_B}). \end{aligned}$$

- Similarly, if caption-specific information of $\mathbf{x}_{\mathcal{C}_B}$ is positive, i.e., if $I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_B} \mid \mathbf{x}_{\mathcal{C}_A}) > 0$:

$$\begin{aligned} \underbrace{I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A} \mid \mathbf{x}_{\mathcal{C}_B}) + I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_B} \mid \mathbf{x}_{\mathcal{C}_A}) + I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A}; \mathbf{x}_{\mathcal{C}_B})}_{(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{OPT}; \mathbf{x}_{\mathcal{C}_A} \mathbf{x}_{\mathcal{C}_B})} &> \underbrace{I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_B} \mid \mathbf{x}_{\mathcal{C}_A}) + I(\mathbf{x}_{\mathcal{I}}; \mathbf{x}_{\mathcal{C}_A}; \mathbf{x}_{\mathcal{C}_B})}_{I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}_A}^{MIN}; \mathbf{x}_{\mathcal{C}_A})} \Rightarrow \\ &\Rightarrow I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{OPT}; \mathbf{x}_{\mathcal{C}_A} \mathbf{x}_{\mathcal{C}_B}) > I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}_A}^{MIN}; \mathbf{x}_{\mathcal{C}_A}). \end{aligned}$$

Therefore, we show that in a setup where a single image is associated with multiple captions, and where at least one caption contains caption-specific information, image representations learned contrastively w.r.t. associated captions contain less information than task-optimal image representation: $\exists i \in \{A, B\}$ such that $I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}}^{OPT}; \mathbf{x}_{\mathcal{C}_A} \mathbf{x}_{\mathcal{C}_B}) > I(\mathbf{z}_{\mathcal{I} \rightarrow \mathcal{C}_i}^{MIN}; \mathbf{x}_{\mathcal{C}_i})$. \square

4.D EXPERIMENTAL SETUP

4.D.1 Datasets

Flickr30k consists of 31,000 images annotated with 5 matching captions (Young et al., 2014).

MS-COCO consists of 123,287 images, each image annotated with 5 matching captions (Lin et al., 2014). The original dataset was introduced for large-scale object recognition.

For both datasets, we use the training, validation, and test splits from (Karpathy and Li, 2015).

4.D.2 Models

We use CLIP and VSE++. Both consist of an image and a text encoder that do not share parameters.

CLIP is a large-scale image-text foundation model (Radford et al., 2021). The model is pre-trained on a collection of 400 million image-text pairs collected from the Web. The encoders are pre-trained using a contrastive loss (InfoNCE) on image-text pairs. The text encoder consists of a 12-layer transformer model, described in (Radford et al., 2019). As for the image encoder, CLIP utilizes various model backbones, such as ResNet (He et al., 2016) and Vision Transformer (Dosovitskiy et al., 2021). In this chapter, we use the ResNet-50 ('RN50') variant of the CLIP image encoder.¹ The CLIP encoders are trained to jointly understand images and text. Therefore, the learned representations generalize to a wide range of different zero-shot (visual) evaluation tasks, such as classification, without task-specific fine-tuning, by using textual prompts.

VSE++ is an image-caption encoder trained from scratch (Faghri et al., 2018). The model features a triplet loss function with a margin parameter $\alpha = 0.2$. The text encoder is a one-layer gated recurrent unit (GRU) (Cho et al., 2014). The available image encoder configurations are ResNet-152 (He et al., 2016) and VGG19 (Simonyan and Zisserman, 2015). In this chapter, we use ResNet-152.

4.D.3 Training

CLIP. To fine-tune CLIP, we follow (Yuksekgonul et al., 2023). All models are fine-tuned for 5 epochs. We employ a cosine-annealing learning rate schedule, with a base learning rate of $2e - 5$, and 100 steps of warm-up. As an optimizer, we use AdamW (Loshchilov and Hutter, 2019) with a gradient clipping value of 2. For the InfoNCE loss, we use the logit-scale (i.e., temperature τ) from the pre-trained CLIP model and fine-tune the logit-scale end-to-end along with the rest of the model parameters.

VSE++. The model is trained for 30 epochs using a linear learning rate schedule with a base learning rate of $2e - 4$. We use the Adam optimizer (Kingma and Ba, 2015) with a gradient clipping value of 2. Instead of the triplet loss, we use the InfoNCE loss similar to (Radford et al., 2021),

For both models, instead of selecting the best-performing model based on the validation set scores, we use the final checkpoint at the end of training.

¹ <https://github.com/openai/CLIP/>

4.D.4 Shortcut Sampling

Our goal is to add the shortcuts in a manner that preserves the original information of the images and captions. For the captions, we append the shortcut at the end of the captions. In order to prevent a tokenizer from tokenizing the shortcut into a single token, we insert spaces between each number of the shortcut. For the images, we place the numbers of the shortcuts at the top of the images, evenly spaced across the entire width of the images (to make sure the shortcut is evenly spaced across the feature map of the image). We always use 6 digits to represent a shortcut. If a shortcut number contains fewer than 6 digits, we fill the remaining positions with zeros for padding. For the MNIST images, we always sample a random image from the set of images representing the number that belongs to (also during evaluation), to prevent overfitting on specific MNIST images. In Figure 4.D.1, we provide four examples of image-caption pairs with randomly added shortcuts. The examples in Figure 4.D.1 show (i) how synthetic shortcuts are added to the image and the caption, and (ii) that the shortcuts preserve the original (task-relevant) information of the images and captions.

4.E OPTIMIZATION OBJECTIVES

4.E.1 InfoNCE

In this chapter, we use InfoNCE loss, $\mathcal{L}_{\text{InfoNCE}}$ (Oord et al., 2018). Given a dual-encoder setup, we optimize a model in two directions: image-to-text (i2t) and text-to-image (t2i). The loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}}^{\text{i2t}} &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{z}_I^i \mathbf{z}_C^i / \tau)}{\exp(\mathbf{z}_I^i \mathbf{z}_C^i / \tau) + \sum_{j \neq i} \exp(\mathbf{z}_I^j \mathbf{z}_C^j / \tau)} \\ \mathcal{L}_{\text{InfoNCE}}^{\text{t2i}} &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathbf{z}_I^i \mathbf{z}_C^i / \tau)}{\exp(\mathbf{z}_I^i \mathbf{z}_C^i / \tau) + \sum_{j \neq i} \exp(\mathbf{z}_I^j \mathbf{z}_C^j / \tau)} \\ \mathcal{L}_{\text{InfoNCE}} &= \frac{1}{2} \mathcal{L}_{\text{InfoNCE}}^{\text{i2t}} + \frac{1}{2} \mathcal{L}_{\text{InfoNCE}}^{\text{t2i}}.\end{aligned}$$

4.E.2 Latent Target Decoding

Latent target decoding (LTD) (Bleeker et al., 2022) is an optimization objective that reduces predictive feature suppression for resource-constrained VL methods. LTD consists of $\mathcal{L}_{\text{InfoNCE}}$ and a reconstruction loss $\mathcal{L}_{\text{recon}}$, which reconstructs the input caption from the latent representation \mathbf{z}_C .



(a) Caption: “A bathroom sink with wood finish cabinets. 0 3 9 9 6 5.”



(b) Caption: “A guy in a brown shirt has just hit a tennis ball. 0 7 7 1 1 4.”



(c) Caption: “A man in shorts is lying on the beach. 0 0 6 9 9 3.”



(d) Caption: “A player up to bat in a baseball game. 1 0 1 9 9 2.”

Figure 4.D.1: Four random samples from the MS-COCO dataset including shortcuts added on both the image and caption.

In the original paper, LTD is implemented in two ways. Firstly, as a dual optimization objective:

$$\mathcal{L}_{\text{InfoNCE+LTD}} = \mathcal{L}_{\text{InfoNCE}} + \beta \mathcal{L}_{\text{recon}}.$$

Secondly, as an optimization constraint in combination with gradient descent by using the method of Lagrange multipliers:

$$\max_{\lambda} \min \mathcal{L}_{\text{InfoNCE+LTD}} = \mathcal{L}_{\text{InfoNCE}} + \lambda \left(\frac{\mathcal{L}_{\text{recon}}}{\eta} - 1 \right).$$

This optimization objective is minimized w.r.t. model parameter, while also being maximized w.r.t. λ . The value of λ is automatically tuned by gradient ascent, such that the reconstruction bound η is met. In this chapter, we use both LTD as a dual optimization objective and an optimization constraint. We select the loss with the highest evaluation scores on the validation set for evaluation.

4.E.3 Implicit Feature Modification

Implicit feature modification (IFM) (Robinson et al., 2021) is a contrastive loss, with an additional perturbation budget ϵ . IFM perturbs the logits value of the similarity scores between the images and captions, such that the model avoids using shortcut solutions for a correct similarity score. IFM subtracts ϵ/τ from the positive logit values and adds ϵ/τ to the negative logits values.

$$\begin{aligned}\mathcal{L}_{\text{IFM}}^{t2i} &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp((\mathbf{z}_I^i \mathbf{z}_C^i - \epsilon)/\tau)}{\exp((\mathbf{z}_I^i \mathbf{z}_C^i - \epsilon)/\tau) + \sum_{j \neq i} \exp((\mathbf{z}_I^j \mathbf{z}_C^i + \epsilon)/\tau)} \\ \mathcal{L}_{\text{IFM}}^{i2t} &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp((\mathbf{z}_I^i \mathbf{z}_C^i - \epsilon)/\tau)}{\exp((\mathbf{z}_I^i \mathbf{z}_C^i - \epsilon)/\tau) + \sum_{j \neq i} \exp((\mathbf{z}_I^j \mathbf{z}_C^i + \epsilon)/\tau)} \\ \mathcal{L}_{\text{IFM}} &= \frac{1}{2} \mathcal{L}_{\text{IFM}}^{t2i} + \frac{1}{2} \mathcal{L}_{\text{IFM}}^{i2t} \\ \mathcal{L}_{\text{InfoNCE+IFM}} &= \frac{1}{2} \mathcal{L}_{\text{IFM}} + \frac{1}{2} \mathcal{L}_{\text{InfoNCE}}.\end{aligned}$$

Similar to (Robinson et al., 2021), we combine IFM and the InfoNCE in a dual optimization objective.

REPRODUCIBILITY

To ensure the reproducibility of the findings presented in this chapter, we have made our code publicly accessible at <https://github.com/MauritsBleeker/svl-framework>.

5

ASSESSING BRITTLENESS OF IMAGE-TEXT RETRIEVAL BENCHMARKS

We continue our investigation by focusing on the topic of evaluating vision-language models (VLMs) in the context of image-text retrieval (ITR) task. Current evaluation methodologies, often rely on coarse-grained datasets where textual descriptions lack the necessary level of detail. This limitation restricts our ability to fully capture the fine-grained relationships between images and text, potentially leading to an underestimation of model brittleness. In other words, models might perform well on current benchmarks but struggle with real-world scenarios where concept granularity is higher.

Motivated by this problem, in this chapter, we address the following research question:

RQ4: How can we improve the evaluation and benchmarking of vision-language models on the image-text retrieval task?

To address this question, we examine the concept of granularity within existing ITR benchmarks and compare them with fine-grained alternatives. We then introduce a novel evaluation framework that incorporates perturbations and a new evaluation metric aimed at capturing semantic similarity and cross-modal relationships. We select four state-of-the-art (SOTA) VLMs and assess their performance using this framework, focusing on robustness to perturbations across both coarse and fine-grained datasets. This chapter contributes to our understanding of how concept granularity affects model performance in the ITR task and suggests potential improvements for evaluation and benchmarking processes.

This chapter is based on the paper “Assessing Brittleness of Image-Text Retrieval Benchmarks from Vision-Language Models Perspective” (Hendriksen et al., 2024).

5.1 INTRODUCTION

Image-text retrieval (ITR) is a bidirectional retrieval task focused on retrieving top- k images or textual captions based on queries in the other modality (Baltrusaitis et al., 2019). This task enhances the connection between visual and textual information, leading to richer and more relevant search results (Cao et al., 2022a). VLMs have achieved state-of-the-art performance in this area (Li et al., 2021a; Chen et al., 2023c; Radford et al., 2021; Xu et al., 2022).

The development of ITR has been significantly supported by open benchmarks, with MS-COCO (Lin et al., 2014; Chen et al., 2015) and Flickr30k (Young et al., 2014) serving as essential evaluation tools. However, we argue that the datasets and evaluation methods for assessing state-of-the-art models require revision due to two key limitations.

Coarse vs. fine-grained datasets. We first focus on the concept granularity of ITR datasets. Here, *granularity* pertains to the specificity of the relationship between images and their textual descriptions (Chen et al., 2023b; Goei et al., 2021; Laenen et al., 2018). Fine-grained datasets provide detailed captions, while coarse-grained datasets offer general descriptions. We contend that key benchmarks like MS-COCO and Flickr30k utilize coarse-grained captions, making it difficult to evaluate models' abilities to identify specific objects or attributes. Some recent work has addressed this by introducing fine-grained dataset augmentations, such as MS-COCO-FG and Flickr30k-FG, which incorporate additional contextual details from images (Chen et al., 2023b).

Robustness. We also examine concept granularity through the lens of model *robustness*. Robustness is vital for VLMs in ITR tasks due to the noise and variability present in real-world data. Common issues include semantic shifts and typographical errors all of which can degrade model performance. Recent research highlights the importance of developing systems that generalize well to out-of-distribution data and resist adversarial attacks (Liu et al., 2024a; Liu et al., 2023; Liu et al., 2024b; Lupart and Clinchant, 2023; Parry et al., 2024; Penha et al., 2022). Furthermore, existing ITR evaluation metrics often rely on binary matches between images and texts, ignoring real-world scenarios where there may be partial semantic overlaps (Messina et al., 2021; Wang et al., 2020; Zhong et al., 2020). Effective evaluation metrics should account for cross-modal relationships to better capture the real-world complexities.

We hypothesize that these limitations contribute to the brittleness of today's ITR evaluation pipeline. To investigate this, we take a two-step approach. First, we assess the granularity of standard ITR benchmarks, MS-COCO and Flickr30k, and their fine-grained counterparts, MS-COCO-FG and Flickr30k-FG. By analyzing features that cap-

ture concept granularity, we can determine how varying levels of descriptive detail influence the performance of VLMs on the ITR task. Second, we extend this analysis to test the robustness of the VLMs using a novel evaluation framework. This framework introduces input perturbations that allow us to test VLMs' sensitivity to word order, redundant information, and lexical variation. We also introduce a cross-modal evaluation metric that extends beyond the traditional binary matching approach to assess the semantic similarities between images and texts. We apply this framework to evaluate the performance of a diverse set of state-of-the-art VLMs – ALIGN (Jia et al., 2021), AltCLIP (Chen et al., 2023c), CLIP (Radford et al., 2021), and GroupViT (Xu et al., 2022). While all models are two-tower architectures trained contrastively, they differ in focus (details are provided in Section 7.4).

In this chapter, we answer the following research questions:

RQ4.1 How does concept granularity – both in textual descriptions and overall dataset composition – impact the performance of VLMs on the ITR task?

RQ4.2 How is the performance of state-of-the-art VLMs (ALIGN, AltCLIP, CLIP, and GroupViT) on the coarse-grained vs. fine-grained datasets impacted by perturbations, particularly in terms of their sensitivity to word order and robustness to variability of user input?

The principal contributions of our research are the following:

- (i) We evaluate the impact of dataset granularity on the performance of vision-language models in ITR using standard benchmarks, MS-COCO and Flickr30k, and their fine-grained counterparts, MS-COCO-FG and Flickr30k-FG.
- (ii) We propose a novel evaluation suite for VLMs on the ITR task, which focuses on the model's compositional understanding and robustness in the context of concept granularity, and features a cross-modal evaluation metric.
- (iii) We evaluate ALIGN, AltCLIP, CLIP, and GroupViT using the proposed framework and find that caption augmentation improves model robustness to perturbations when evaluated on the ITR tasks. Notably, we observed the most substantial decline in model sensitivity to word order, contrasting with previous findings in the domain. This highlights the necessity for more benchmarks that capture both coarse and fine-grained semantic relationships between images and text.

5.2 PRELIMINARIES

Notation. We follow notation from prior work (Bleeker et al., 2024; Brown et al., 2020). Let \mathcal{D} be a dataset of N image-text tuples: $\mathcal{D} = \{(\mathbf{x}_{\mathcal{I}}^i, \{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k)\}_{i=1}^N$. Each tuple $i \in N$ consists of a single image $\mathbf{x}_{\mathcal{I}}^i$ and k corresponding texts (captions) $\mathbf{x}_{\mathcal{C}_j}^i$, where $1 \leq j \leq k$. All texts are considered relevant to the image $\mathbf{x}_{\mathcal{I}}^i$. We derive sets of queries \mathcal{Q} and candidates \mathcal{C} from the dataset \mathcal{D} . Let $\mathcal{Q}_{\mathcal{T}}$ represent the set of text queries, where $\mathcal{Q}_{\mathcal{T}} \subseteq \mathcal{Q}$. Let $\mathcal{Q}_{\mathcal{I}}$ represent the set of image queries, where $\mathcal{Q}_{\mathcal{I}} \subseteq \mathcal{Q}$. Similarly, $\mathcal{C}_{\mathcal{T}} \subseteq \mathcal{C}$ and $\mathcal{C}_{\mathcal{I}} \subseteq \mathcal{C}$ represent the sets of text and image candidates respectively. Let $q \in \mathcal{Q}$ and $c \in \mathcal{C}$ represent a query and a candidate item respectively.

A query q may originate from either the text modality $q \in \mathcal{Q}_{\mathcal{T}}$ or the image modality $q \in \mathcal{Q}_{\mathcal{I}}$, while a candidate c may similarly originate from either the text modality $c \in \mathcal{C}_{\mathcal{T}}$ or the image modality $c \in \mathcal{C}_{\mathcal{I}}$. Let $E_{\theta_1} : \mathcal{Q} \rightarrow \mathbb{R}^d$ be the encoder function mapping textual queries $q \in \mathcal{Q}_{\mathcal{T}}$ to d -dimensional vectors: $\mathbf{q} = E_{\theta_1}(q)$. Similarly, we write $E_{\theta_2} : \mathcal{C} \rightarrow \mathbb{R}^d$ for the encoder function mapping image queries $c \in \mathcal{C}_{\mathcal{I}}$ to d -dimensional vectors: $\mathbf{c} = E_{\theta_2}(c)$.

Let $f_{\text{rel}} : \mathcal{Q} \times \mathcal{C} \rightarrow \mathbb{R}$ be a relevance function that computes the relevance of a query-candidate pair. We write $f_S : \mathcal{Q} \times \mathcal{C} \rightarrow \mathbb{R}$ for a scoring function that takes a query and a candidate, maps them into d -dimensional space, normalizes the vectors so that they lie on d -dimensional hypersphere and computes their similarity. Finally, $f_{\text{sim}} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ denotes a similarity function that computes a similarity score between the two d -dimensional vectors. We assume that all vectors lie on the surface of a d -dimensional hypersphere. Formally, this implies that $\|q\| = \|c\| = 1$ where $\|\cdot\|$ denotes the Euclidean norm.

Task. We focus on the task of *cross-modal retrieval*, which involves matching queries in one modality (e.g., text or image) to candidates in a different modality.

The retrieval process can occur in two ways: (i) *text-to-image retrieval* (t2i): given a textual query $q \in \mathcal{Q}_{\mathcal{T}}$ and a set of candidate images $\mathcal{C}_{\mathcal{I}}$, rank the images by their relevance to q ; (ii) *image-to-text retrieval* (i2t): given an image query $q \in \mathcal{Q}_{\mathcal{I}}$ and a set of text candidates $\mathcal{C}_{\mathcal{T}}$, rank the texts by their relevance to q . For both tasks, dedicated encoders are used to map images and texts into a shared d -dimensional representation space. Once encoded, we compute the similarity between the query and candidate in this shared space to derive relevance scores.

Performance is typically evaluated using Recall@k (R@k), where $k = \{1, 5, 10\}$, and the sum of recall (rsum).

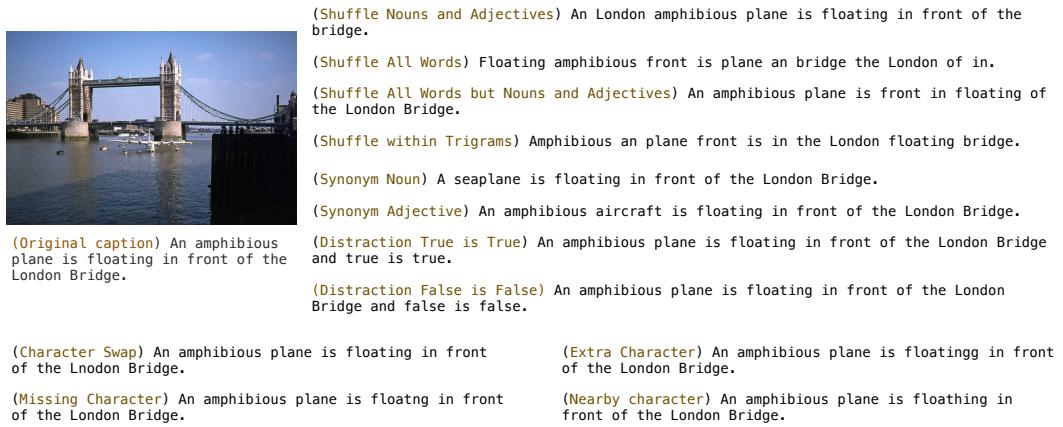


Figure 5.1: Overview of selected perturbations with examples.

5.3 CONCEPT GRANULARITY IN IMAGE-TEXT RETRIEVAL DATASETS

In this section, we outline the features for analyzing the granularity of concepts in ITR datasets. We will also describe the datasets selected for evaluation and perform an analysis based on the provided definition of granularity.

5.3.1 Granularity Features in Image-Text Retrieval

We focus on features that contribute to defining the granularity of ITR datasets, focusing on noun phrase (NP) and caption-level characteristics.

NP-level granularity. This section discusses linguistic features contributing to NP granularity in captions.

Modifiers of the noun. Adjectives and complement phrases (CPs) provide details about objects in images (Pesahov et al., 2023; Zhao et al., 2022). By quantifying these modifiers, we assess the detail and granularity associated with objects (Li et al., 2022d). Specifically, we count the number of adjectives and CPs per identified noun in captions.

Semantics: Concept depth. Concept depth reflects the semantic understanding captured within individual concepts in captions, indicating a deeper comprehension of the depicted scene (Xu et al., 2023). Datasets with deeper conceptual information offer more detailed descriptions of visual content (Piasecki et al., 2009). We measure concept depth by calculating the minimum depth of the corresponding synsets, considering the maximum depth across all synsets associated with a word.

Determiners: Articles, quantifiers. The use of articles and quantifiers impacts the specificity of noun descriptions (Jurafsky and Martin, 2009). Analyzing their occurrences offers insights into the explicitness and precision of noun specifications. We quantify the occurrences of articles and quantifiers linked to identified nouns in captions.

Table 5.1: Granularity vs. Coarseness in ITR Datasets.

Level	Features	MS-COCO	MS-COCO-FG	Flickr30k	Flickr30k-FG
NP	Adjectives	0.76	1.05	1.14	1.3
	CPs	1.56	1.99	1.81	2.19
	Articles	2.14	2.34	2.27	2.55
	Quantifiers	0.12	0.13	0.26	0.27
	Concept Depth	7.89	7.91	7.97	7.97
Caption	Caption Length	52.39	56.38	63.61	68.29
	Words per Caption	10.59	11.48	12.34	13.67
	Concept Diversity	9.14	10.04	9.86	10.68

Caption-level granularity. Next, we consider caption-level features.

Caption length. The character count of a caption indicates the amount of information conveyed (Lewis and Frank, 2016). Longer captions are likely to include more details, contributing to finer granularity. We measure the total word count for each caption.

Number of words. The total word count is indicative of caption richness (Lewis and Frank, 2016). A higher word count suggests a more elaborate description, signaling finer granularity. We count the total number of words in each caption.

Semantic diversity of concepts per caption. Concept diversity is essential for analyzing granularity within ITR datasets (Jurafsky and Martin, 2009). It reflects the range of ideas and semantic complexity captured in a caption. We compute the ratio of unique synonyms to the total word count in each caption.

5.3.2 Granularity Analysis

Next, we analyze the selected datasets in terms of granularity versus coarseness, with a focus on various linguistic aspects at both the NP and caption levels.

Datasets

In this chapter, we use the following datasets:

- (i) **MS-COCO** (Lin et al., 2014), a large-scale object detection, segmentation, and captioning dataset that consists of 123,287 images and 616,435 captions; each image is annotated with 5 captions.
- (ii) **Flickr30k** (Young et al., 2014), an image caption corpus consisting of 158,915 crowd-sourced captions describing 31,783 images; each image is annotated with 5 captions.

- (iii) **MS-COCO-FG and Flickr30k-FG** (Chen et al., 2023b), augmentation of Flickr30k and MS-COCO, respectively, with captions that contain additional contextual details extracted from the associated images.

For all datasets, we use the training, validation, and test splits from (Karpathy and Li, 2015).

Table 5.1 presents the results of analyzing our datasets in terms of granularity. For Flickr30k vs. Flickr30k-FG, we observe a 21% increase in the number of concept phrases in the extended dataset. This indicates a richer description of scenes with additional details. The concept depth remains unchanged. While the fine-grained dataset offers more detailed descriptions, the semantic complexity of the concepts remains largely unchanged. Similarly, we note a 38% increase in the number of adjectives per caption in MS-COCO-FG over MS-COCO. This suggests a more descriptive and nuanced portrayal of visual content. The concept depth exhibits only a marginal increase, implying that the semantic understanding of concepts is slightly enhanced in the fine-grained version. Overall, the fine-grained datasets demonstrate higher scores across features than their standard counterparts. Thus, they offer more detailed and descriptive captions, amounting to improved granularity.

5.4 EVALUATION FRAMEWORK

5.4.1 Perturbations

We introduce several perturbations to evaluate the robustness and performance of VLMs in the context of ITR. These perturbations focus on word order sensitivity and robustness to noise in input. We draw inspiration from prior work that highlights the limitations of large language models in processing word order (Hessel and Schofield, 2021; O’Connor and Andreas, 2021; Pham et al., 2021; Yuksekgonul et al., 2023) and handling noisy input (Thomas and Kovashka, 2020; Jin et al., 2020; Fan et al., 2021; Zhuang and Zuccon, 2022; Wang et al., 2022c).

Word-level Perturbations

Word-level perturbations are applied at the level of individual words within a caption. The focus is on investigating the model’s robustness to typos and synonyms. The perturbations types include:

Typos. Typos are common in real-world scenarios, and evaluating a model’s response to such errors is crucial for ensuring its practical usability in information retrieval (IR) (Sidiropoulos and Kanoulas, 2022; Zhuang and Zuccon, 2022) and on the

image-caption retrieval (ICR) task in particular (Wang et al., 2022c). This perturbation assesses the model’s ability to handle input variations introduced by typographical mistakes, providing insights into its robustness in retrieving images given textual descriptions. Typos perturbations aim to assess the model’s resilience to typographical errors. This type has been previously tested on sentiment analysis, duplicate question detection, and natural language inference (Wang et al., 2021a; Li et al., 2018). However, it has not been applied in the context of evaluating VLMs on the ITR task. The subtypes are as follows.

- **Character Swap:** Swaps two random adjacent word characters in a caption, simulating the introduction of a typo through character transposition. This perturbation allows us to evaluate the model’s ability to recognize and correct character-level errors.
- **Missing Character:** Removes a randomly selected character from the input text, mimicking the effect of a typo where a character is omitted. This perturbation tests the model’s robustness in understanding and completing partial textual information.
- **Extra Character:** Adds an extra random character to the input text, simulating the insertion of a typo. This perturbation assesses the model’s ability to handle additional characters and maintain accurate image-caption associations despite minor textual variations.
- **Nearby Character:** Replaces a character in the input text with a nearby character on the keyboard, emulating the introduction of a typo due to the proximity of keys. This perturbation explores the model’s sensitivity to keyboard-related errors.

Synonyms

Synonym-based perturbations aim to assess the model’s adaptability and robustness to variations in language, specifically focusing on the substitution of nouns and adjectives with their synonyms. This perturbation type is motivated by the need to evaluate VLMs capacity to comprehend and retrieve images and captions when faced with lexical variations that convey similar meanings (Jin et al., 2020; Fan et al., 2021). Specifically, we focus on testing the models’ capacity to retrieve the right image using semantically similar nouns and adjectives. The subtypes are as follows.

- **Synonym Noun:** This perturbation involves replacing k nouns in a given caption with their synonyms. The motivation behind this perturbation is to examine how well the model handles variations in nouns, which is important for accurate and descriptive image-caption associations.

- **Synonym Adjective:** This perturbation implies replacing k adjectives in a given caption with their synonyms. Adjectives play a vital role in expressing characteristics and qualities associated with visual elements in an image. Introducing synonym substitutions in adjectives aims to assess the model’s proficiency in maintaining the descriptive quality of captions when faced with lexical variations.

Sentence-level Perturbations

Sentence-level perturbations are applied at the level of sentences in a caption. The focus is on evaluating the model’s resilience to distracting elements, compositionality-related challenges, and sensitivity to word order.

Distraction-Based Perturbations. Distraction-based perturbations aim to evaluate the model’s robustness to distracting elements within captions. Specifically, we focus on the statements that are always true and do not add any meaningful content to the caption. The motivation is to understand how well the model can filter out relevant information from distractors, a critical skill for accurate image-caption retrieval in the presence of additional context (Thomas and Kovashka, 2020).

- **Distraction True is True:** This subtype appends to caption distracting statement “true is true.” It evaluates the model’s handling of additional distracting information that is semantically coherent but not directly related to the original content.
- **Distraction False is False:** This subtype appends to caption distracting statement “false is false,” assessing the model’s resilience to distracting information.

Compositionality-Related Perturbations

Compositionality-related perturbations assess the model’s ability in the context of compositionality (Partee, 1995; Yuksekgonul et al., 2023), focusing on its sensitivity to word order changes within sentences.

Sensitivity to Word Order. This category of perturbations tests the model’s sensitivity to word order changes within sentences.

- **Shuffle Nouns and Adjectives:** This subtype involves shuffling the order of nouns and adjectives within the input sentence. The motivation is to examine how well the model can handle changes in the arrangement of descriptive elements, crucial for capturing the visual details of an image accurately.
- **Shuffle All Words:** Randomly shuffling the order of all words in the input sentence to assess the model’s general flexibility in understanding and generating coherent captions despite drastic changes in word order. This perturbation aims to reveal the model’s adaptability to varied sentence structures.

- **Shuffle All Words But Nouns and Adjectives:** Shuffling all words except for nouns and adjectives tests the model’s ability to maintain the key descriptive elements in their original positions, examining its proficiency in preserving the essential details while undergoing significant rearrangement. In practice, it implies keeping the nouns and adjectives in fixed positions and randomly shuffling all the other words.
- **Shuffle within Trigrams:** Dividing the input sentence into trigrams and shuffling the order of words within each trigram evaluates the model’s response to localized word rearrangements. This perturbation offers insights into the model’s sensitivity to changes in smaller, contextually relevant segments of the sentence.
- **Shuffle Trigrams:** Dividing the input sentence into trigrams and shuffling the order of entire trigrams assesses the model’s ability to comprehend and generate captions when faced with larger-scale rearrangements. This perturbation provides a broader perspective on the model’s understanding of sentence composition and structure in diverse contexts.

5.4.2 Evaluation Metric

The current evaluation framework for ITR faces challenges due to the binary match assumption, the focus on intra-modality comparisons, and the disregard of cross-modal relationships across image-caption tuples (Kaur et al., 2021; Carrara et al., 2018; Messina et al., 2021; Wang et al., 2020; Jiang et al., 2017; Jiang et al., 2017). Such limitations hinder the comprehensive assessment of model performance, failing to capture the relationships between visual and textual content. To address these shortcomings, we propose a novel evaluation metric that uses similarity functions to estimate relevance scores across modalities and image-caption tuples. Our goal is to evaluate not only explicit matches but also the overall relevance between queries and candidates, even when explicit labels are unavailable. To achieve this, we define a metric based on both *perfect match* cases and *cross-modal relevance*.

We operate in a setup when, given a query q , and a ranked list of top- k retrieved results $K = [c^1, \dots, c^k]$, we want to obtain a list of the relevance scores $[rel^1, \dots, rel^k]$ where rel^i denotes the relevance for the i -th retrieved candidate.

Perfect Match. When explicit matching labels are available, we assign a relevance score of 1 to perfect matches. This applies to both text-to-image and image-to-text retrieval:

- (i) *Text-to-Image Retrieval (t2i):* The retrieved image $c \in \mathcal{C}_{\mathcal{T}}$ is considered a perfect match if it is the ground-truth image for query $q \in \mathcal{Q}_{\mathcal{T}}$:

$$f_{rel}(q, c) = 1 \quad \text{if} \quad \exists i \in N \text{ such that } q \in \{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k \wedge c = \mathbf{x}_{\mathcal{T}}^i.$$

- (ii) *Image-to-Text Retrieval (i2t)*: The retrieved caption $c \in \mathcal{C}_{\mathcal{T}}$ is considered a perfect match if it is the ground-truth caption for query $q \in \mathcal{Q}_{\mathcal{I}}$:

$$f_{rel}(q, c) = 1 \quad \text{if } \exists i \in N \text{ such that } q = \mathbf{x}_{\mathcal{I}}^i \wedge c \in \{\mathbf{x}_{\mathcal{C}_j}^i\}_{j=1}^k.$$

Cross-Modal Relevance. When explicit labels are unavailable (i.e., no perfect matches exist), the relevance score is computed based on the similarity between the encoded query and candidate vectors. This approach allows us to measure how well the model aligns cross-modal pairs (text and images) in the shared representation space. The scoring function f_S is defined as:

$$f_S(q, c, E_{\theta_1}, E_{\theta_2}) = \begin{cases} f_{sim}(E_{\theta_1}(q), E_{\theta_2}(c)) & \text{if } q \in \mathcal{Q}_{\mathcal{T}} \text{ and } c \in \mathcal{C}_{\mathcal{I}}, \\ f_{sim}(E_{\theta_2}(q), E_{\theta_1}(c)) & \text{if } q \in \mathcal{Q}_{\mathcal{I}} \text{ and } c \in \mathcal{C}_{\mathcal{T}}. \end{cases} \quad (5.1)$$

We use cosine similarity as the similarity function: $f_{sim}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|}$.

Overall Metric. To evaluate model performance across ranked results, we measure relevance while considering the rank position of the results:

$$DCG_{CM}^p = \sum_{i=1}^p \frac{\text{rel}^i}{\log_2(i+1)}, \quad (5.2)$$

where p represents the rank position up to which the score is computed.

5.5 EXPERIMENTS

5.5.1 Models

For our experiments, we select four VLM that demonstrate SOTA performance across vision-language (VL) tasks, with a specific emphasis on ITR. All selected models are dual-encoder networks trained using contrastive learning on image-text data; however, they embody a diverse array of methodologies, thereby ensuring a comprehensive evaluation.

ALIGN (Li et al., 2021a) is a VLMs that addresses the challenge of costly curation processes in VL representation learning by leveraging a noisy dataset of over one billion image alt-text pairs from the Conceptual Captions dataset. Employing a simple contrastive dual-encoder architecture, ALIGN learns to align visual and language representations effectively. The model achieves SOTA results on a variety of VL tasks, outperforming more complex cross-attention models. The learned representations enable zero-shot image classification and support cross-modality search with complex text and image queries, showcasing the effectiveness and scalability of the ALIGN

model in large-scale VL tasks.

AltCLIP (Chen et al., 2023c) is a multilingual VLM built upon CLIP (Radford et al., 2021). It enhances CLIP’s capabilities by incorporating a pre-trained multilingual text encoder XLMR and employing a two-stage training schema. In the first stage, knowledge distillation from CLIP is conducted through teacher learning, followed by contrastive learning in the second stage, where the model is trained on a small set of Chinese and English text-image pairs. AltCLIP achieves SOTA performances on a variety of VL tasks. Furthermore, AltCLIP closely matches CLIP’s performance, indicating that simple alterations to CLIP’s text encoder can lead to extended capabilities in handling multilingual tasks.

CLIP (Radford et al., 2021) is a dual encoder pre-trained on a dataset of 400 million (image, text) pairs collected from the internet. Its pre-training enables zero-shot transfer to downstream tasks, where natural language references learned visual concepts or describes new ones. Benchmarked across over 30 diverse computer vision datasets, including OCR, action recognition, and fine-grained object classification, CLIP demonstrates remarkable versatility and competitiveness, often matching or surpassing fully supervised baselines without requiring task-specific training. Similar to the GPT family, CLIP exhibits proficiency across a wide range of tasks during pre-training, showcasing its potential as an efficient and effective method for large-scale VL representation learning and ITR.

GroupViT (Xu et al., 2022) reintroduces the grouping mechanism of grouping semantic regions into deep networks, enabling the automatic emergence of semantic segments. Trained contrastively on a large-scale paired image-text dataset, GroupViT learns to group image regions into progressively larger arbitrary-shaped segments. This hierarchical approach, facilitated by the flexibility of the global self-attention mechanism in the transformer architecture, allows GroupViT to dynamically form different visual segments for various input images, each representing a distinct semantic concept.

5.5.2 Experiments Overview

To answer our RQs, we run the following experiments:

Set 1: Coarse vs Fine-Grained Datasets Evaluation across Selected Models (RQ4.1). In these experiment, we evaluate the impact of concept granularity in both textual descriptions and dataset composition on VLMs performance in the ITR task. We validate our evaluation framework by comparing our results to those reported in a previous study (Chen et al., 2023b). This study is relevant because it critiques current ITR benchmarks and proposes enhanced evaluations for fine-grained cross-modal se-

mantic matching. Moreover, Chen et al. (2023b) introduced augmented benchmarks (MS-COCO-FG and Flickr30K-FG) that we incorporate into our experiments. We run the ITR task on both standard image-caption datasets (MS-COCO and Flickr30k) and their more fine-grained counterparts (MS-COCO-FG and Flickr30K-FG). The models are evaluated on image-to-text (i2t) and text-to-image (t2i) tasks, and we report the recall at 1 for both. This experiment allows us to assess how refining textual descriptions and increasing dataset granularity impact model performance.

Set 2: Model Robustness and Perturbation Sensitivity (RQ4.2). In these experiments, we the robustness to perturbations of state-of-the-art VLMs (ALIGN, AltCLIP, CLIP, and GroupViT) on the coarse-grained vs. fine-grained datasets. We apply 13 perturbations across the four selected datasets (MS-COCO vs. MS-COCO-FG, and Flickr30k vs. Flickr30K-FG). The perturbations are designed to test the models' sensitivity to changes in word order and robustness to noisy input. We analyze the performance drop of the models after each perturbation and measure their sensitivity to word order, lexical variations, and typos.

5.5.3 Results

Set 1: Coarse vs. Fine-Grained Datasets Evaluation across Selected Models (RQ4.1). To address RQ4.1, we evaluate models R@1 performance for both i2t and t2i retrieval and compare the results between the original datasets (MS-COCO, Flickr30k) and their fine-grained versions (MS-COCO-FG, Flickr30k-FG). Table 5.2 highlights that refining the captions improves performance in most cases. Across datasets, we observe significant improvements in R@1 scores. The highest performance gain is a 29.11% improvement in CLIP for t2i retrieval on the Flickr30k dataset. On average, scores increase by 12.63% on MS-COCO and 10.05% on Flickr30k. Specifically, MS-COCO exhibits an 8.14% increase for i2t retrieval and a 17.11% increase for t2i, while Flickr30k shows a 4.75% rise in i2t scores and a 15.35% rise for t2i. However, there are exceptions, particularly in the CLIP MS-COCO t2i and GroupViT MS-COCO i2t tasks, where refined captions do not improve the scores. Despite these few exceptions, the overall results demonstrate that refining textual descriptions enhances retrieval performance, with the greatest benefits observed in t2i retrieval, which saw an average 16.23% improvement compared to a 6.44% increase in i2t retrieval. Therefore, we answer RQ4.1 as follows: the results suggest that concept granularity in captions positively impacts the performance of VLMs on ITR tasks, especially for text-to-image retrieval.

Set 2: Model Robustness and Perturbation Sensitivity (RQ4.2). To address RQ4.2, we assess the robustness of four VLMs (ALIGN, AltCLIP, CLIP, GroupViT) to various perturbations across MS-COCO, Flickr30k, and their refined counterparts. We apply

Table 5.2: Model performance, on the i2t and t2i tasks. “DCG” is short for “ DCG_{CM} .“ “MS-...” is short for “MS-COCO” and “Fli...” for Flickr30k.”

Model	i2t				t2i				rsum		
	R@1	R@5	R@10	DCG	R@1	R@5	R@10	DCG	i2t	t2i	
MS-COCO	ALIGN	42.22	54.42	60.48	2.45	22.93	42.15	51.01	1.60	157.12	116.09
	AltCLIP	40.95	53.44	58.64	2.43	22.47	41.85	50.90	1.61	153.03	115.22
	CLIP	33.66	45.29	50.08	2.32	16.15	33.11	42.06	1.66	129.03	91.32
	GroupViT	24.88	34.38	35.72	1.97	8.29	18.90	25.59	1.41	94.98	52.78
MS-COCO-FG	ALIGN	44.59	56.55	64.20	2.50	25.60	45.64	54.65	1.61	165.34	125.89
	AltCLIP	43.97	57.23	61.83	2.51	25.45	45.86	54.75	1.63	163.03	126.06
	CLIP	38.16	50.38	55.20	2.43	16.15	33.11	42.01	1.66	143.74	91.27
	GroupViT	24.88	34.38	35.72	1.97	9.58	21.38	28.68	1.42	94.98	59.64
Flickr30k	ALIGN	70.52	83.58	88.90	3.03	35.56	58.78	67.64	1.70	243.00	161.98
	AltCLIP	67.98	82.46	86.40	2.99	33.06	56.42	65.74	1.69	236.84	155.22
	CLIP	58.06	72.54	79.30	2.85	19.30	39.74	49.22	1.70	209.90	108.26
	GroupViT	35.34	49.24	50.80	2.20	8.36	19.26	26.02	1.38	135.38	53.64
Flickr30k-FG	ALIGN	75.28	87.38	90.80	3.10	39.80	64.76	73.44	1.73	253.46	178.00
	AltCLIP	71.66	85.96	87.40	3.05	37.10	61.02	70.60	1.72	245.02	168.72
	CLIP	63.70	77.72	82.60	2.95	24.92	46.00	55.60	1.73	224.02	126.52
	GroupViT	38.50	53.88	52.30	2.26	8.92	20.98	28.54	1.38	144.68	58.44

the proposed perturbations to contrast how well models handle changes in word order, lexical variations, and typos, in the coarse-grained vs. fine-grained settings. The results are shown in Table 5.2. The results indicate consistent drops across most perturbation-dataset pairs. The most notable decrease is caused by the *shuffle all words* perturbation, where randomly shuffling all words in captions leads to the largest score drops, underscoring the models’ reliance on correct word order for accurate retrieval. In contrast, the *lexical variation* perturbation has the smallest effect, indicating a greater model resilience to synonym substitution. Interestingly, while most perturbations negatively affect performance, in some cases, refined datasets exhibit better robustness. For example, on MS-COCO-FG, models show smaller relative performance drops for when compared to MS-COCO. This trend is less consistent for Flickr30k-FG, which shows smaller performance drops than Flickr30k for only 5 of the 13 perturbations. This discrepancy may be due to the inherently more detailed nature of Flickr30k captions, making additional granularity less beneficial than in MS-COCO, which has coarser captions. Overall, these findings highlight the sensitivity of VLMs to perturbations, with word order being particularly critical. Interestingly, this contradicts prior work

Table 5.3: Rsum after applying perturbation.

Perturbation	MS-COCO	MS-COCO-FG	Flickr30k	Flickr30k-FG
ALIGN				
No perturbation	116.09	125.89	161.98	168.72
Shuffle N&A	100.00	109.58	139.33	145.39
Shuffle all words	85.78	97.58	120.39	130.77
Shuffle all but N&A	98.03	116.59	133.67	154.19
Shuffle within trigrams	101.70	116.12	144.65	154.16
Shuffle trigrams	104.23	117.86	145.06	156.83
Distraction	112.17	124.91	156.20	163.51
Lexical variation	108.88	119.46	157.79	161.61
Typos	103.07	115.25	152.83	152.01
AltCLIP				
No perturbation	115.22	126.06	155.22	178.00
Shuffle N&A	96.84	107.54	133.63	154.82
Shuffle all words	88.41	98.91	121.62	132.39
Shuffle all but N&A	100.08	113.69	135.68	159.44
Shuffle within trigrams	101.60	113.66	138.82	160.87
Shuffle trigrams	103.81	115.35	143.14	163.60
Distraction	110.20	120.63	157.08	173.07
Lexical variation	107.46	118.20	148.64	174.12
Typos	100.91	112.60	141.32	161.71
CLIP				
No perturbation	91.32	91.27	108.26	126.52
Shuffle N&A	31.23	72.24	86.06	99.74
Shuffle all words	41.24	60.87	69.19	77.82
Shuffle all but N&A	28.93	75.40	82.52	99.31
Shuffle within trigrams	26.11	74.12	84.57	100.26
Shuffle trigrams	30.60	76.41	91.08	103.33
Distraction	84.05	89.93	105.75	121.10
Lexical variation	74.12	84.04	101.26	139.32
Typos	66.30	76.86	87.99	105.37
GroupViT				
No perturbation	52.78	59.64	53.64	58.44
Shuffle N&A	43.62	49.00	46.82	49.87
Shuffle all words	41.94	46.82	47.83	46.89
Shuffle all but N&A	49.08	54.58	51.82	48.32
Shuffle within trigrams	48.18	54.52	51.72	54.36
Shuffle trigrams	48.56	53.98	52.84	47.52
Distraction	51.18	58.23	53.47	59.91
Lexical variation	48.61	53.71	49.78	53.89
Typos	43.11	49.81	47.65	50.44

on this topic where authors argue that reshuffling word order does not affect ITR performance (Yuksekgonul et al., 2023). Therefore, we answer RQ4.2 by stating that fine-grainedness of a dataset positively impacts the performance of VLMs on ITR task.

5.5.4 Model Input Analysis

We further investigate the robustness of the models under different types of caption perturbations. For each model, we collect a sets of collect perturbed captions and their corresponding rsums. We categorize all the perturbed captions into three groups based on their impact on model performance: (i) perturbed captions causing performance decrease, (ii) perturbed captions causing performance increase, and (iii) perturbed captions with no change in performance. We proceed by calculating Jaccard similarity for each category (increased, decreased, no change) across all image-caption pairs within a dataset. This analysis helps identify patterns in how each model’s performance is affected by different perturbations.

The results are shown in Table 5.4. The highest Jaccard similarity scores are observed for perturbed captions that do not impact the models’ performance. This indicates that certain types of captions consistently lead to outcomes where the model’s performance remains unaffected, regardless of perturbation type. It implies that the models exhibit a degree of robustness towards specific types of caption variations, which indicates a level of generalizability. Conversely, the lowest Jaccard similarity scores are associated with perturbed captions that increase models performance. This indicates that captions that lead to outcomes where the model’s performance improves vary significantly.

5.6 RELATED WORK

5.6.1 Cross-Modal Retrieval

Cross-modal retrieval (CMR) methods learn a latent space, where the similarity of concepts from different modalities can be measured using a distance metric such as cosine or Euclidean distance. Some of the earliest approaches in CMR used canonical correlation analysis (Gong et al., 2014; Klein et al., 2014). This was later followed by the emergence of a dual encoder architecture that combined recurrent and convolutional components, gaining prominence in the field and often employing a hinge loss (Frome et al., 2013; Wang et al., 2016b). Further advancements have increased effectiveness through techniques like hard-negative mining (Faghri et al., 2018). Subse-

Table 5.4: Jaccard similarity across perturbations, averaged per model.

Jaccard similarity per group			
	decreased	increased	unchanged
MS-COCO			
ALIGN	0.1680	0.0802	0.6810
AltCLIP	0.1676	0.0714	0.6911
CLIP	0.1792	0.0700	0.7025
GroupViT	0.1653	0.0658	0.7215
MS-COCO-FG			
ALIGN	0.1640	0.0902	0.6613
AltCLIP	0.1659	0.0725	0.6752
CLIP	0.1778	0.0740	0.6805
GroupViT	0.1598	0.0650	0.7073
Flickr30k			
ALIGN	0.1342	0.0513	0.6136
AltCLIP	0.1650	0.0838	0.6429
CLIP	0.1879	0.0858	0.6510
GroupViT	0.1646	0.0761	0.6913
Flickr30k-FG			
ALIGN	0.1275	0.0518	0.6071
AltCLIP	0.1615	0.0807	0.6425
CLIP	0.1868	0.0799	0.6249
GroupViT	0.1542	0.0728	0.6890

quently, the incorporation of attention mechanisms, such as dual attention (Nam et al., 2017), stacked cross-attention (Lee et al., 2018), and bidirectional focal attention (Liu et al., 2019), further improved performance. Other work aims to improve CMR performance through modality-specific graphs (Wang et al., 2021b), or image and text generation modules (Gu et al., 2018), or learning sparse multimodal representations (Nguyen et al., 2024). And there is domain-specific research focusing on CMR in various fields such as fashion (Laenen et al., 2018; Goei et al., 2021), e-commerce (Hendriksen et al., 2022), cultural heritage (Sheng et al., 2021b), and cooking (Wang et al., 2021b).

Recent methods use transformer-based dual encoders trained on extensive data. AL-BEF (Li et al., 2021a) aligns unimodal representations before fusion, X-VLM (Zeng et al., 2022) adds a cross-modal encoder for fine-grained VL representations. Florence (Yuan et al., 2021) uses adaptation models for object-level representations, and CLIP

(Radford et al., 2021) predicts image-caption pairs. ALIGN (Li et al., 2021a) uses a dual encoder on image alt-text pairs. FILIP (Yao et al., 2022) features late multimodal interaction, and SLIP (Mu et al., 2022) combines language and image self-supervision. DeCLIP (Li et al., 2022c) improves CLIP pretraining via self-supervision and cross-modal supervision. AltCLIP (Chen et al., 2023c) uses a pre-trained multilingual text encoder and a two-stage training schema. GroupViT (Xu et al., 2022) reintroduces the grouping mechanism to vision transformers, dynamically forming visual segments for various images.

Another line of work adopts transformer encoders (Vaswani et al., 2017) for the ITR task (Messina et al., 2021), adapting models like BERT (Devlin et al., 2019). ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) introduce a two-stream architecture, while B2T2 (Alberti et al., 2019), VisualBERT (Li et al., 2019b), Unicoder-VL (Li et al., 2020a), VL-BERT (Su et al., 2020), and UNITER (Chen et al., 2020b) propose single-stream architectures. Oscar (Li et al., 2020d) incorporates caption object tags with region features, and BEIT-3 (Wang et al., 2022b) adapts multiway transformers. This chapter focuses on transformer-based dual encoder models due to their performance on various VL tasks. We select four SOTA methods and provide a comparative analysis of their performance on the ITR task.

5.6.2 Vision-Language Model Evaluation

The evaluation of VLMs assesses their performance across various tasks and datasets. Standard benchmarks are MS-COCO (Lin et al., 2014; Chen et al., 2015) and Flickr30k (Young et al., 2014) for tasks like image captioning, visual question answering, and ITR. More fine-grained benchmarks like MS-COCO-FG and Flickr30k-FG (Chen et al., 2023b) are due to limitations in concept granularity and diversity. Specialized datasets like CUB-200 (Welinder et al., 2010), ABO (Collins et al., 2022), and Fashion200k (Han et al., 2017) cater to specific domains. Large-scale and domain-specific datasets like Conceptual Captions (Sharma et al., 2018), XMarket (Bonab et al., 2021), and Recipe1M (Marin et al., 2021) enable evaluation of VLMs in real-world applications.

Evaluating the robustness and generalization of VLMs is key for understanding real-world performance. Studies have explored VLMs robustness to adversarial attacks (Zhao et al., 2023b), domain shifts, and input perturbations (Yuksekgonul et al., 2023), aiming to identify vulnerabilities and improve robustness. Adversarial attacks on VLMs have been extensively studied in visual question answering (Bartolo et al., 2021; Cao et al., 2022b; Kaushik et al., 2021; Kovatchev et al., 2022; Li et al., 2021b; Sheng et al., 2021a; Wallace et al., 2019; Xu et al., 2018; Zhang et al., 2022a) and image captioning (Aafaq et al., 2021; Chen et al., 2018; Xu et al., 2019).

Another important aspect of model evaluation is metrics. For CMR tasks, the quality of retrieved top-k images or texts given a query in a different modality is a primary focus. Common metrics include Recall@K (Kaur et al., 2021; Hendriksen et al., 2023), adaptations of Discounted Cumulative Gain (Carrara et al., 2018), Normalized Discounted Cumulative Gain (Messina et al., 2021), Precision-Recall curves (Wang et al., 2020; Zhong et al., 2020; Xu et al., 2020), F-score (Jiang et al., 2017), Mean Average Precision (Wang et al., 2014), and Mean Reciprocal Rank (Qin et al., 2016; Jiang et al., 2017).

Unlike prior work in this domain, we focus on both benchmark performance and robustness analysis, while incorporating a diverse set of evaluation metrics to provide a comprehensive understanding of VLM capabilities in the context of the ITR task.

5.7 CONCLUSION

In this chapter, we address the brittleness of the evaluation pipeline on the ITR task, emphasizing two primary concerns: the coarseness of existing benchmarks and the limitations of current evaluation metrics. Through our analysis, we compare standard datasets, MS-COCO and Flickr30k, with their fine-grained counterparts, MS-COCO-FG and Flickr30k-FG. We propose an evaluation framework that encompasses a taxonomy of perturbations and a new evaluation metric designed to improve the robustness of ITR assessments.

We selected four state-of-the-art VLMs – AltCLIP, ALIGN, CLIP, and GroupViT – for our experiments and evaluate their performance on the ITR task using the novel framework. We discover that caption augmentation improves the performance of VLMs on the ITR tasks. We observe the biggest decline when testing models sensitivity to word order which is opposite to prior findings in this domain. Specifically, the fine-grainedness of a dataset positively impacts VLMs performance on the ITR task, with finer-grained datasets, MS-COCO-FG and Flickr30k-FG, leading to higher performance for all selected models. This finding highlights the need for more benchmarks that would capture both coarse- and fine-grained semantic relationships between images and text. Therefore, we suggest that future benchmarking efforts should focus on developing rich datasets that enable experimenters to systematically vary the level of granularity to better evaluate the capabilities of VLMs.

However, this study has limitations. We focused on a specific set of perturbations and datasets, which may not encompass the full spectrum of real-world scenarios. Additionally, while we selected leading models in the domain of ITR, evaluating a broader range of VLMs could yield a more comprehensive understanding of their performance across diverse datasets and evaluation frameworks. Expanding our eval-

uation to include models with varied architectures and training methodologies could provide deeper insights into their robustness and generalization in the ITR task.

Future work should aim to extend our framework by incorporating additional perturbations and datasets, as well as expanding the range of evaluated models. Another promising avenue includes exploring other facets of VLM performance on the ITR task, such as interpretability and domain adaptation, to further improve our understanding of their capabilities and limitations.

As a result, our answer to RQ4 is that improving the evaluation and benchmarking process of vision-language models on the image-text retrieval task involves addressing the granularity of benchmarks and limitations of current evaluation metrics. Finer-grained datasets improve model performance even when input variations are introduced, highlighting the sensitivity of models to changes in input data. An evaluation framework incorporating a taxonomy of perturbations can test model robustness, emphasizing the need for detailed datasets and robust evaluation methods to accurately assess model capabilities and develop models resilient to input variability.

6

PREDICTING PURCHASE INTENT FOR PRODUCT RETRIEVAL

Next, we switch back to the topic of product retrieval and focus on the problem of understanding and predicting purchase intent in e-commerce. Prior work in this area has primarily focused on user sessions where the customer is identified by the platform. However, in practice, a significant portion of online shopping journeys begin anonymously.

The context of cross-device interactions adds another layer of complexity. Modern consumers frequently switch between devices such as smartphones, tablets, and computers during their purchasing journey. Each device provides a distinct set of modalities, including screen size, input method, and usage context, all of which can influence user behavior and intent.

This discrepancy presents an interesting challenge and an opportunity to explore how purchase intent can be effectively predicted in both identified and anonymous sessions in a cross-device scenario.

Therefore, motivated by the need to better understand and predict purchase intent in this multifaceted environment, we pose the following research question:

RQ5: How can we facilitate product retrieval by predicting purchase intent in cross-device setting?

To answer this question, we will sample session logs from the e-commerce platform to identify signals indicative of purchase intent. These signals include session duration, dwell time, device type, channel, and search queries. By engineering features based on these insights, we aim to develop predictive models tailored for both anonymous and identified sessions. We proceed by analyzing the performance of these models.

This chapter was published at the 44th European Conference on Information Retrieval (SIGIR eCom 2020) under the title “Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customer” (Hendriksen et al., 2020).

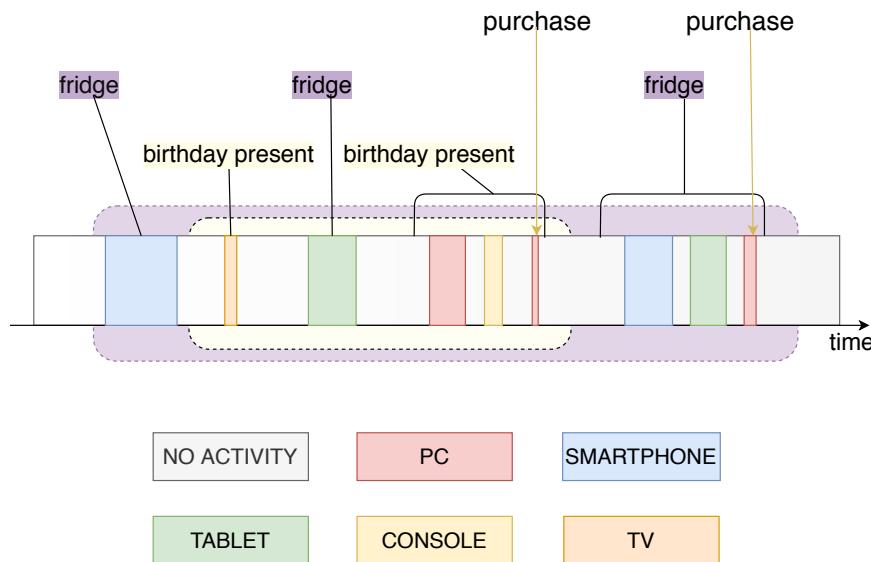


Figure 6.1: Customer journeys across sessions, with multiple interests and devices; the colors indicate different devices.

Through this chapter, we aim to improve our understanding of user behavior across devices in e-commerce settings.

6.1 INTRODUCTION

Information retrieval (IR) technology is at the heart of today's e-commerce platforms, in the form of search engines, recommenders, and conversational assistants that connect users to the products they may be interested in (Rowley, 2000). To help improve the effectiveness of IR technology in an e-commerce context, the problem of analyzing, modeling, and, ultimately, predicting customers' purchase intent has been studied extensively in academia and industry (Bellman et al., 1999; Agichtein et al., 2006; Lo et al., 2016).

Purchase intent prediction. Here, purchase intent is defined as a predictive measure of subsequent purchasing behavior (Morwitz and Schmittlein, 1992).

Figure 6.1 illustrates the complexities of customer behavior during a sequence of sessions, when multiple tasks, interests, and devices may play a role. Areas in the back of the figure are meant to signify different user journeys across time, purple for one that is focused on fridges, yellow for one that is focused on a birthday present. Colored rectangular blocks in the front indicate different devices used by the user. Initial exploration of a relatively expensive item (a fridge) starts on a smartphone and continues on a tablet, while the journey ends with a purchase of a fridge on a PC. The purchase of a fridge is interleaved with the purchase of a (lower-priced) birthday

present, with initial exploration on a PC, followed by further exploration on a TV and PC, and, ultimately, a purchase on a PC.

Online search behavior that targets transactions has been analyzed at scale at least since the work by Broder (2002), who identified a class of so-called *transactional* queries, where the user is seeking to reach a page with more interaction opportunities, e.g., to conduct a purchase, download or sign-up. In particular, factors influencing online purchases have been described as early as in 2002 (George, 2002), and work on predicting purchases goes back to at least the work of (Ben-Shimon et al., 2015), where the task was to predict whether a given customer is going to purchase within a given session.

Challenges. Despite the many advances, purchase intent prediction still has many challenges (Tsagkias et al., 2020). In particular, previous work on purchase intent prediction has focused mostly on customers of an e-commerce platform who are identified or recognized by the platform. A diverse range of models has been considered, from traditional feature-based models such as boosted decision trees to sequence-based neural models such as RNNs. However, based on the analysis of de-identified data from an e-commerce website available to us, more than 50% of traffic comes from anonymous users. Purchase intent detection for anonymous users is particularly challenging because it cannot rely on historical information about the user on which many of the existing models rely.

Features for purchase intent prediction. In this chapter, we focus on identifying signals that suggest purchase intent in an anonymous and identified setting. We do this by analyzing purchase vs. non-purchase sessions sampled from a large European e-commerce website and testing the features based on our observations on a production-ready model. We further test the obtained feature sets on five other classifiers to explore the generalizability of our findings. In particular, we include features derived from session-based data such as page dwell time and customer-specific data such as the number of days since the last purchase. Session-based features have the advantage that they are available both during sessions when a user is identified (i.e., the customer has logged-in or is recognized through cookies) and anonymous sessions (when the customer is not known). Customer-related features are only available during identified sessions. Interestingly, many of the features proposed previously (Seippel, 2018) apply only to identified sessions: *purchase intent prediction for anonymous sessions has been studied very little.*

To fill this gap, we analyze a dataset of more than 95 million sessions, sampled from four weeks of anonymized user interaction data in a European e-commerce platform. We answer the following research questions:

RQ5.1: *How do purchase sessions differ from non-purchase sessions?* In Section 6.4 we compare purchase vs. non-purchase sessions in such aspects as session length, temporal variations, device and channel type, queries. Among others, we find out that

purchase sessions tend to be longer than non-purchase ones, customers are more likely to purchase in the evening and during a weekday, and more likely to own more than 1 device.

RQ5.2: *What are the important session-based features that allow us to tell purchase sessions apart from non-purchase sessions? What are the important historical features that should inform predictors for identified sessions? How does the importance of features change across the session?* Based on the experiments described in Section 6.5, we conclude that historical features related to previous purchasing behavior are highly important for detecting purchases in the identified setting. For the anonymous setting, however, dynamic features related to page dwell time and sequence of pages are most important. Besides, the importance of dynamic features increases as the session continues, while the importance of static features decreases.

RQ5.3: *How effective are models used for purchase intent prediction for anonymous vs. identified sessions? Furthermore, to which degree do the proposed features help improve performance for anonymous sessions?* In Section 6.5, we show that in the anonymous setting, tree-based and neural classifiers demonstrate the best performance, and adding extra features to models improves F_1 by about 17%. In contrast, for the identified setting all models demonstrate high performance and adding extra features do not provide a significant gain.

The principal contributions of our research are the following:

- We conduct an in-depth analysis of a real-world customer interaction dataset with more than 95 million sessions, sampled from a European e-commerce platform. We identify session features such as device type and conversion rate, weekday, channel type, and features based on historic customer data such as number of previous orders and number of devices to distinguish between purchase and non-purchase sessions (see Section 6.4).
- We define two feature sets for purchase prediction, tailored towards anonymous sessions and identified sessions (see Section 6.5).
- We extend an existing production-ready model to evaluate our proposed features and run additional experiments with classifiers generally used for this task. We find F_1 improvements of up to 17% in purchase intent prediction for anonymous sessions and reach an F_1 of 96% for identified sessions on held-out data collected from a real-world retail platform (see Section 6.5).

6.2 BACKGROUND AND DEFINITIONS

In our study, we operate with the following definitions.

A *session* is a sequence of requests made by a single end-user during a visit to

a particular site. A session ends if the user is idle for more than 30 minutes. We define two types of sessions: *purchase sessions*, during which the customer buys an item, and *non-purchase sessions*, during which the customer does not buy anything. In connection to this, we define *purchasers* as customers who had at least one purchase session, whereas *non-purchasers* are customers who were identified but have never purchased anything. We furthermore distinguish between *identified sessions*, where a customer is logged in or recognized with a browser cookie, and *anonymous sessions* where this is not the case. Additionally, we denote the number of actions taken during a given session as the *session length*, where an *action* corresponds to opening a new web page, submitting a search query, or adding/removing an item to/from the shopping basket.

Device switch is the act of changing the type of browsing device between two consecutive sessions that belong to the same journey. For instance, if a customer first explores the platform on a smartphone and afterward accesses the platform on a PC, she switches from a smartphone to a PC.

A *channel* indicates the way through which a customer enters the platform. For example, if the customer comes to the platform via an advertisement, she uses a paid channel.

The *conversion rate* denotes the fraction of visits during which a purchase was made (Moe and Fader, 2004). We use this metric to compare device popularity in a purchasing context. We calculate the conversion rate by dividing the number of purchasing sessions by the overall number of sessions. In order to protect sensitive information, we only report *standardized conversion rates* for each device; since we are interested in differences across devices types, this suffices for our purposes. The standardized conversion rate is computed by subtracting the mean conversion rate per device type from the desired conversion rate and dividing the result by the standard deviation of the device-specific conversion rate. For instance, if our device specific conversion rates are *Conversion Rates* = {0.5, 0.2, 0.3}, the mean of device-specific conversion rates is $\overline{\text{Conversion Rate}} = 0.33$ and the standard deviation of conversion rates is $\sigma = 0.12$. Therefore, the resulting standardized conversion rates are *Standardized Conversion Rates* = {1.34, -1.07, -0.27}

6.3 DATASET DESCRIPTION

In this section, we describe how we extract a dataset consisting of anonymized user interaction data from the search logs of an e-commerce platform, and summarize dataset statistics.

Data Collection. Our dataset comprises four weeks (28 days) of anonymized visits

Table 6.1: Dataset statistics.

Description	Total
Sessions	94,402,590
Anonymous	55,305,709
Logged-in or recognized	39,096,881
Logged-in or recognized customers	6,125,781
Queries	31,185,176
Device types	PC, Smartphone, Tablet, Game Console, TV

sampled from a European e-commerce platform in October 2019. The original sample of the log entries includes a unique non-personal customer identifier (for identified users), the type of browsing device used during the session, as well as a timestamp for every query, and a URL of each clicked page. We convert all the timestamps to the Central European Time Zone (CET). We additionally recorded the price of every product the customers have seen and the prices of the items they ended up buying. In cases where a customer starts a session without logging in and ends up logging in at a later point in the session, we assign the session to the customer.

To filter out bot traffic, we apply several measures related to location and device type (Bomhardt et al., 2005). First, we filter out sessions based on location, to only include entries from the European countries from which the majority of the customers come; bots come mostly from non-European IPs, especially North-America. Second, we specify the set of device types we are interested in and remove all the entries from other devices, leaving us with *PC*, *Smartphone*, *Tablet*, *Game Console*, and *TV*; bots often do not specify a device type.

Dataset Statistics. Table 6.1 provides descriptive statistics of the resulting dataset. The dataset contains 95,757,177 sessions, out of which 54,144,152 (about 56.5%) are anonymous. In total, the dataset contains 9,663,509 identified users. We additionally keep track of the device types used for browsing and distinguish between five such device types: PC, smartphone, tablet, game console, and TV. The table also lists the number of search queries; these are the queries submitted during the sessions captured in the log.

6.4 CHARACTERIZING PURCHASE INTENT

We explore customer behavior and, in particular, the difference in the behavior of purchasing and non-purchasing users. These explorations aim to identify characteristics

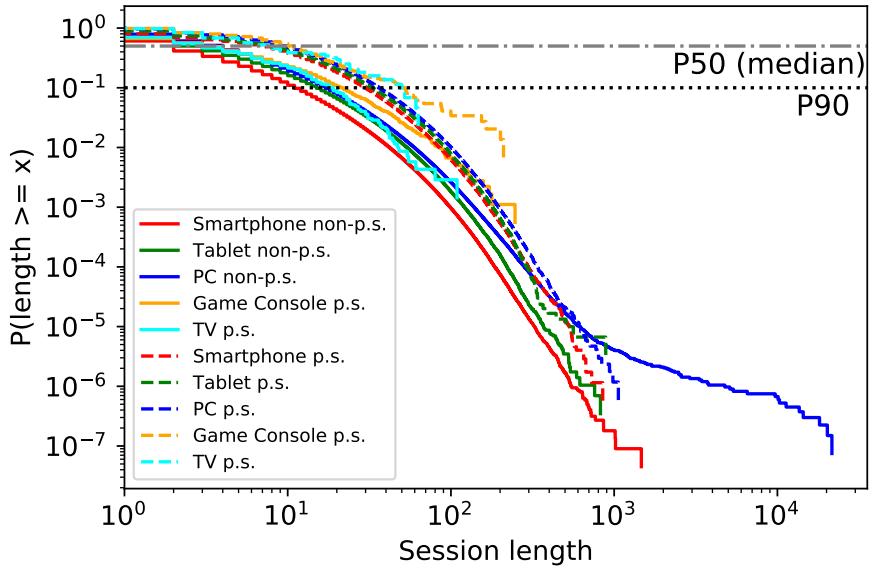


Figure 6.2: CCDF of the session length per device type for purchase sessions (p.s.) and non-purchase sessions (non-p.s.).

that may help us improve the effectiveness of purchase intent predictors. We analyze several aspects of sessions, such as the length of purchase and non-purchase sessions, the temporal characteristics of sessions, and device information. Furthermore, we investigate the channels from which customers start sessions and issue queries during purchase sessions and non-purchase sessions.

6.4.1 Session Length

First, we examine the overall session length for purchase sessions and non-purchase sessions. Figure 6.2 plots the complementary cumulative distribution function (CCDF) of the session lengths of purchase sessions and non-purchase sessions per device type.

As can be seen in the area between the P50 and P90 percentiles in Figure 6.2, purchase sessions are in general longer than non-purchase sessions. Moreover, the purchase session length per device varies less than the non-purchase session length per device. It can be explained by the fact that non-purchase sessions can be both very short or rather long, depending on the underlying user intents. For instance, a user could quickly look something up or spend some time exploring the catalog. On the other hand, in the case of purchase sessions, user intentions are less ambiguous. Usually, users look for a specific product that they have in mind and, upon finding it, proceed to purchase.

From a device perspective, the shortest sessions take place on smartphones, whereas sessions on tablets are generally longer. The longest sessions occur on the PC, TV, and game console. This finding holds for both purchase sessions and non-purchase

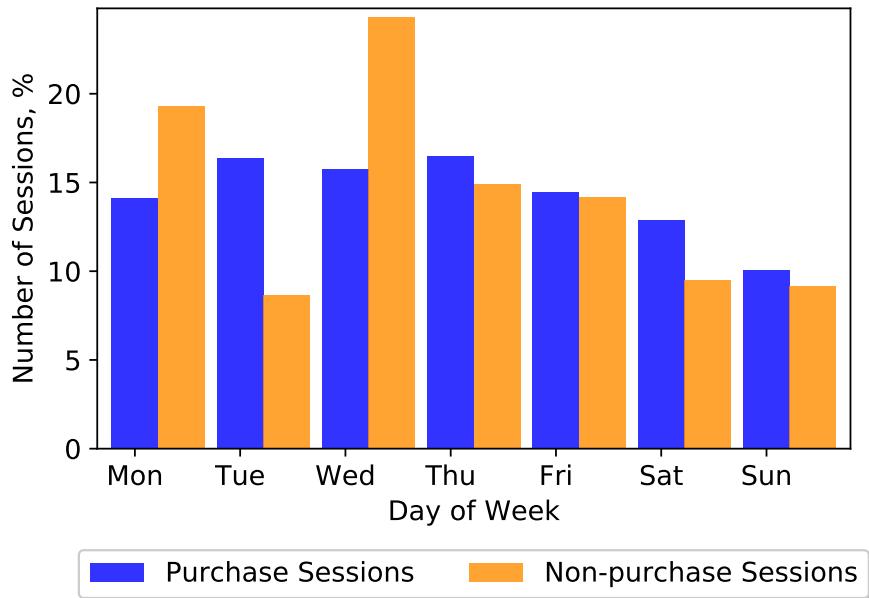


Figure 6.3: The fraction of purchase sessions and non-purchase sessions across days of the week w.r.t. total amount of purchase and non-purchase sessions. Most activity occurs on weekdays.

sessions. However, in the tail of the distributions, the distinction between purchase session length and non-purchase session length is not as clear as between the P50 and P90 percentiles. The non-purchase session length distribution on the PC has an exceptionally long tail. Overall, we can attribute these findings to the fact that smartphones have a smaller screen and are therefore less convenient for longer sessions. Tablet screens are bigger than smartphone screens; hence, the sessions can last longer. The PC screen is the biggest one, and therefore PC users exhibit even longer sessions.

6.4.2 Temporal Variations

Next, we look into the temporal characteristics of purchase sessions and non-purchase sessions, such as their distribution across days of the week and the sessions' starting hours.

First, we want to understand customer activity during the days of the week. Figure 6.3 shows how the number of purchase and non-purchase sessions varies across days of the week. The three most popular days for purchase sessions are Thursday, Tuesday, and Wednesday. In total, the purchase sessions of these three days amount to 48.55% of all purchase sessions. On the other hand, the least popular purchase days are Sunday, Saturday, Monday, and Friday. They contribute to 51.45% of purchase sessions. The observed pattern of purchase behavior hints at the fact that customers prefer to buy during weekdays, which aligns with their workweek. Besides, we conclude that the lower purchase activity on Monday and Friday attributes to their proximity to

weekends.

In the case of non-purchase sessions, the most active session days are Wednesday, Monday, and Thursday. Altogether, these days contribute to 58.55% of non-purchase sessions. The least active days are Tuesday, Sunday, Saturday, and Friday. All the sessions of these days amount to 41.44%. Just like for purchase sessions, the activity for non-purchase sessions also centers around weekdays. However, the difference between the three most active days and the four least active days for non-purchase sessions is bigger than the corresponding difference for purchase sessions. For purchase sessions, the difference is only 2.9%, whereas, for non-purchase sessions, the difference is 17.11%. Moreover, Tuesday, the 2-nd most popular day for purchase sessions is the least popular day for non-purchase sessions. On the other hand, Monday, the 2-nd most popular day for non-purchase sessions is the 3-rd least popular day for purchases. The observation indicates that people need time to consider a purchase before making the buying decision. Hence, they spend Monday, the first day of the new week on considering the purchase, and the purchase itself happens on Tuesday or later in the week. In general, the most active day of the week is Wednesday, whereas the least active day is Sunday. These findings strongly suggest that user behavior depends on the day of the week. In general, people are most active on weekdays, during their workweek, their activity peaks in the middle of the week. On the other hand, at the beginning and end of the workweek, user activity is generally lower.

Next, we look at user behavior on the level of the hour during which a session starts. As mentioned in Section 7.4.1, all the hours are represented in CET. Figure 6.4 shows how purchase sessions and non-purchase sessions spread across the hours of the day.

As expected, the least active hours are in the early morning, in the period from 1 am to 3 am. That can be explained by the fact that most people sleep during the night. (Note that the majority of the e-commerce platform customers come from Europe; hence, time does not vary that much.) Moreover, the activity on the platform during the period from 10 am till 5 pm is stable both for purchase and non-purchase sessions, whereas the most active hours are in the evening, i.e., from 6 pm till 8 pm. In general, our observations correspond to the established rhythm of the daily life of the majority of people, who sleep during the night, browse e-commerce platforms both during work hours and in the evening after work.

6.4.3 Channel Types

Next, we look at whether channel types distributions change across purchase and non-purchase sessions. We define the following channel types: *direct* where a user enters the platform directly; *paid* where a user enters the platform through search engine advertisement, and *organic* where a user enters the platform through a web search

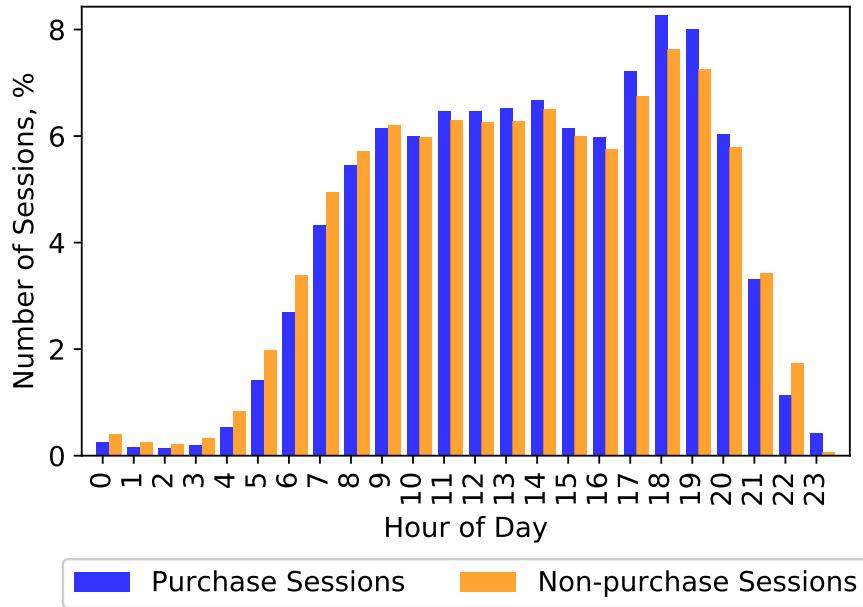


Figure 6.4: The fraction of purchase session and non-purchase sessions across the hours of the day w.r.t. total amount of purchase and non-purchase sessions. Most purchase sessions start in the evening.

Table 6.2: Channel types for purchase and non-purchase sessions.

Channel	Sessions		Stand conv. rate
	Purchase (%)	Non-purchase (%)	
Direct	71.07	77.30	-0.56
Paid	16.74	12.92	0.54
Organic	11.78	7.83	0.94
Other	0.31	1.05	-1.33

engine and unpaid results. Table 6.2 displays the channel distribution across purchase and non-purchase sessions.

Both for purchase and non-purchase sessions, the direct channel is the most used channel to enter the platform. However, for purchase sessions, the percentage of sessions which start with the direct channel is 8.06% less than the fraction of non-purchase sessions, which started with the direct channel. The second most popular channel for purchase and non-purchase sessions is a paid channel. However, in the case of this channel, the fraction of purchase sessions is 12.92% bigger than the corresponding channel type fraction for non-purchase sessions. The organic channel is the third channel in terms of popularity for both session groups. The organic channel fraction for purchase sessions is 50.40% bigger for purchase sessions when compared with non-purchase sessions.

Overall, during purchase sessions, users are more likely to enter the platform via paid or organic channels, whereas for non-purchase sessions the direct channel is

Table 6.3: User device statistics per session.

Device(s)	Purchasers (%)	Non-purchasers (%)
> 1 device	24.05	16.22
1 device	75.95	83.78
2 devices	22.23	15.39
3 devices	1.82	0.82
4 devices	0.01	0.01
5 devices	0	0

more common. It can be explained by the fact that purchasers decide to converge after being offered an advertisement or a search result that matches their interest, whereas non-purchasers may enter the platform directly to explore the catalog.

6.4.4 Devices

In this subsection, we investigate purchase intent from the perspective of device types. In particular, we look at the device types used by purchasers and non-purchasers and analyze device switches.

Device type

First, we want to understand how many users are using multiple devices and which devices customers use for purchase and non-purchase sessions.

Table 6.3 shows how many devices purchasers and non-purchasers own. The majority of users from both groups are single-device users. However, the fraction of single-device purchasers is 9.35% smaller than the corresponding fraction of non-purchasers. On the other hand, the fraction of multi-device users for purchasers is 45.28% bigger than the corresponding fraction for non-purchasers. In general, multi-device users represent almost a quarter of the purchasers. As the number of devices increases, the difference between purchasers and non-purchasers grows. Our observations support the statement that multi-device users tend to be more engaged (Montanez et al., 2014).

Next, we examine the distributions of purchase and non-purchase sessions across device types and device-specific standardized conversion rates; see Table 6.4. The PC is the device with the highest conversion rate. Indeed, the fraction of purchase sessions is 30.70% bigger than the fraction of non-purchase sessions. The device with the second-highest conversion rate is a tablet. For this device, purchase sessions are 7.12% more frequent than non-purchase sessions. The Smartphone is the device with the second-lowest conversion rate. For this device, the number of purchase sessions is

Table 6.4: Purchase and non-purchase sessions per device type and standardized conversion rates.

Device	Purchase sessions (%)	Non-purchase sessions (%)	Stand. conv. rate
Smartphone	47.00	58.09	-0.56
PC	44.97	34.40	1.61
Tablet	8.03	7.50	0.61
Game Console	0.01	0.01	-0.40
TV	0.01	0.01	-1.25

19.10% less frequent than the number of non-purchase sessions.

Game consoles and TVs are relatively new devices in e-commerce; hence, sessions with these devices are relatively less frequent. Nevertheless, based on our observations, we find that the game console is a device with the third-highest conversion rate. Interestingly, its conversion rate is close to that of the smartphone. It can be explained by the fact that device functionalities of smartphones and tablets in e-commerce context blur due to the similarity of their interfaces and screen sizes. The number of purchase sessions on a game console is 15.47% less than the number of non-purchase sessions. The TV is the least common device, with the lowest conversion rate. The number of purchase sessions on this device is 34.01 less than the number of non-purchase sessions.

We can explain our findings by the fact that customers use different devices for different purposes. For example, PCs and tablets seem to be used for the purchase, whereas smartphones, game consoles, and TVs for exploration.

Device switches

Next, we analyze how users switch between devices before a purchase session.

Figure 6.5 shows device transition probability, including self-transitions. Generally, the situation when a user remains on the same device is the most likely outcome for all devices, except TV. There, the self-transition probability is lower than the probability of switching from TV to PC, a device with the highest self-transition probability. A probability of remaining on a smartphone is 5.03% lower than a self-transition probability for PC, whereas a probability to remain on a tablet is 17.28% lower than the probability to remain on PC. The game console has the second-lowest self-transition probability.

Next, we consider connections between two different devices. We characterize those interconnections based on how likely a user is to switch from one device to another one and vice versa.

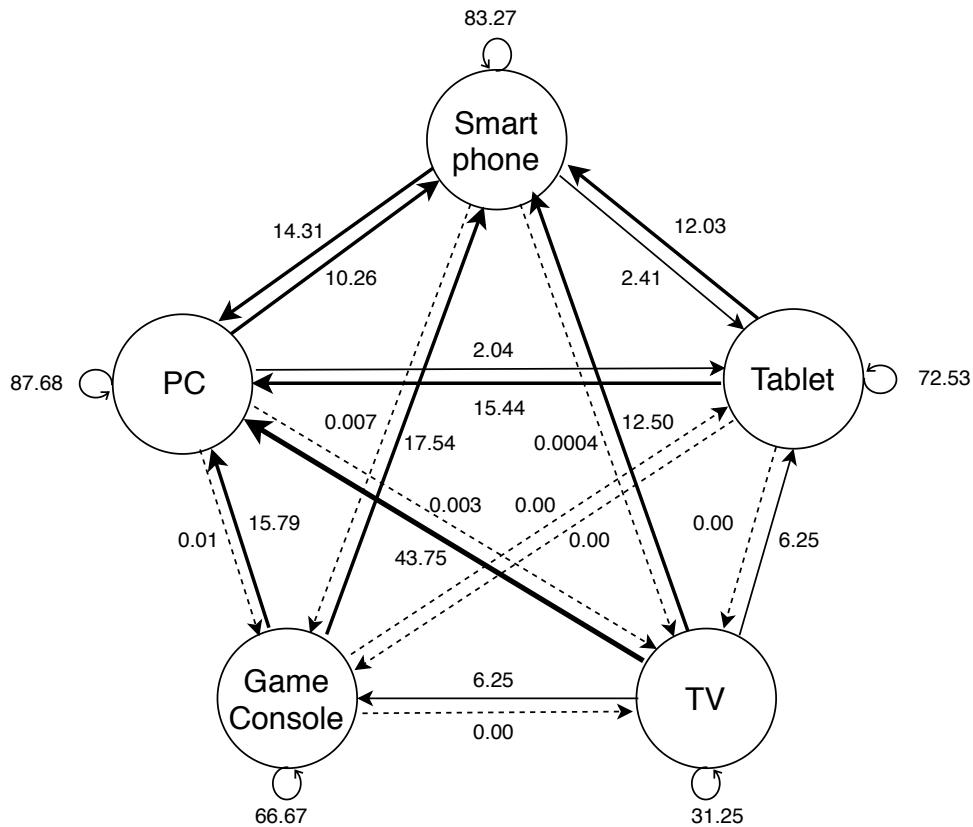


Figure 6.5: Device transition probability before a purchase session, including self-transitions. The thickness of an arrow indicates the connection strength; the dashed line is the weakest connection.

Strong interconnections Some pairs of devices have high probability interconnections. The strongest connection is between a smartphone and a PC, the two most popular devices. The second strongest connection is between PC and tablet. There is a bigger discrepancy between probability rates, with the probability of switching to PC being 656.86% higher than of switching to tablet. The third strongest interconnection is between a smartphone and a tablet with a stronger connection switch to a smartphone, a more popular device. The probability of switching to a smartphone is 399.17% higher than switching to a tablet. Overall, the three interconnections form a triangle that includes the three most popular devices: PC, smartphone, and tablet.

One-sided interconnections For a one-sided interconnection there is a high probability of switching from one device to another, but a close to zero probability of switching back. There are six cases of this type in Figure 6.5. TV is the device with the largest number of one-sided interconnections, with PC, smartphone, tablet, and game console. In all cases, the transition probability is low when TV is a target device, which can be explained by relative difficulty to purchase on TV. The strongest one-sided interconnection is between TV and PC. The probability of switching from TV to PC is 43.75%, the highest transition probability for TV, and the highest probability to transition to

Table 6.5: Queries per Session for Purchase and Non-Purchase Sessions per Device Type

Device	Purchase Sessions		Non-Purchase Sessions	
	Query/Session	%	Query/Session	%
Smartphone	4	52.92	0.05	42.40
PC	2	36.38	0.09	57.54
Tablet	4	10.67	0.01	0.05
Game Console	0.01	0.01	0.01	0.01
TV	0.01	0.01	0.01	0.01
Avg	3.16	100	0.06	100

another device. We explain this by the fact that PC is one of the most popular devices for purchase. The second most likely device people switch to from TV is a smartphone, whereas a probability to switch to a tablet or game console is 6.25%.

Another device with a significant number of one-sided interconnections is the game console. Apart from the connection with TV discussed above, the device also has this connection type with smartphone and PC. The transition probability is close to zero when a game console is a target device. Unlike the situation with TV, game console has a higher probability of switching to a smartphone, whereas the probability of switching to PC is 1.75% less.

In general, all one-sided interconnection cases include switching from a less common device type such as game console or TV to a more conventional device, such as PC, smartphone, or tablet.

Weak interconnections In some cases, the switch between two devices rarely happens, i.e., the transition probability is close to zero. As can be seen in Figure 6.5, there is only one case of this type. It is a connection between a game console and a tablet.

Overall, the analysis of device switches before a purchase session supports the conclusion that users tend to switch from less popular devices such as TV and game console to more popular ones such as PC, smartphone, and tablet.

6.4.5 Queries

The next aspect of purchase intent that we examine is queries. We look at the number of queries in purchase and non-purchase sessions and per device type. In total, the dataset contains 31,185,176 queries, 1,302,195 or 4.17% of which are unique. Given the number of sessions in the dataset, we can conclude that queries are infrequent.

Table 6.5 shows the query per session frequencies across five devices for both purchase and non-purchase sessions. Besides, it also demonstrates which devices are most

Table 6.6: Unique query counts for purchase and non-purchase sessions per device type. The percentage is computed w.r.t. total number of queries per purchase or non-purchase session.

Device	Purchase sessions		Non-purchase sessions	
	Count	%	Count	%
Smartphone	180,542	36.89	321,640	39.57
PC	224,763	45.92	312,855	38.49
Tablet	83,881	17.14	175,909	21.64
Game Console	136	0.03	1,841	0.23
TV	46	0.01	582	0.07
Total	489,368	100.00	812,827	100.00

popular for querying during purchase and non-purchase sessions. Overall, queries are more common in purchase sessions. This can be explained by the fact that querying is more likely to happen when customers are determined to buy something.

Naturally, queries are most common for smartphones, PCs and tablets, and uncommon for game consoles and TVs. Indeed, the current interface of game console and TV makes it difficult to type queries, especially when compared to a PC or a smartphone.

The PC has the highest query per session frequency for non-purchase sessions and second-highest frequency for purchase sessions. A smartphone has the second-highest query per session frequency for non-purchase sessions and the highest query frequency per non-purchase session. Tablet, on the contrary, has the third-highest frequency for non-purchase sessions and the highest frequency for purchase sessions.

When it comes to query distributions per device for purchase and non-purchase sessions, the ranking is somewhat consistent for both groups. During purchase sessions, most queries are issued on a smartphone, whereas during non-purchase sessions PC prevails. On the other hand, PC is the second most popular device for purchase sessions, whereas for non-purchase sessions smartphone takes the second place. Tablet is third for both groups. In general, the query distribution across devices correlates with the session distribution across devices (see Table 6.4).

Next, we look at the number of unique queries for purchase and non-purchase sessions and per device. Table 6.6 shows unique queries count and their corresponding fractions. The fractions are computed w.r.t. the total number of queries per session type and device. Overall, during purchase sessions users issue less unique queries, it holds for every device class but a PC. This can be explained by the fact that during purchase sessions users may retype a previous query to revisit the results they have seen earlier, whereas non-purchasers want to explore and hence use more unique queries.

6.4.6 Purchase Intent Characteristics

What have we learned from the log analysis conducted in this section that might help us to devise better models for purchase intent prediction? We found out that purchase sessions tend to be longer what suggests that session length is an essential indicator of purchase intent. Besides, the difference in session length depends on the type of device customer uses. Moreover, we discovered how the day of week and hour of the day influence purchase behavior. In particular, customers are more likely to buy during the weekdays and in the evening. From the perspective of channels, there is a difference, too. In particular, for non-purchase sessions, the direct channel is more common, whereas purchase sessions are more likely to start with paid or organic channels. From the device perspective, we found out that multi-device users are more common among purchasers. Besides, we figured the probability of purchase for every device and characterized transitions between devices. After looking into queries in the dataset, we discovered that during purchase sessions, users issue more queries per session. Besides, during purchase session, there are less unique queries.

6.5 PREDICTING PURCHASE INTENT

Next, we turn to predict purchase intent when a user is anonymous (“anonymous setting”) and when a user is logged-in or recognized (“identified setting”). The goal of our experiments is to evaluate how the features which we discovered during dataset exploration influence purchase predictor performance in both settings. To accomplish this, we derive a feature set for each setting, and evaluate the features by adding them to an existing production-ready model, based on a Random Forest. To showcase the generalizability of our findings, we additionally test the impact of our features on five additional popular classifiers. To investigate how the models’ ability to predict purchase evolves throughout a session, we evaluate all models on 11 session steps (corresponding to the visits of 10 pages). We are interested in longer sessions because the outcome of such sessions is more difficult to predict. As we do not want to evaluate the model’s performance on the very last step, (where the outcome is clear), we set up a buffer of 2 pages. Therefore, we filter out all the sessions which are shorter than 12 pages. We conclude the section by analyzing the features which contributed most to the model performance in both the anonymous and the identified setting, and explore how dynamic and static feature importance change as the session continues.

Table 6.7: Complete feature set. “Dynamic” indicates that a feature may change during a session.

	Feature	Dynamic	Baseline
Session	current page dwell time, mean	✓	✓
	current page dwell time, σ	✓	✓
	page sequence score	✓	✓
	number of pages	✓	✓
	channel type		
	start hour		
	week day		
	device type		
History	device conversion rate		
	number of orders		✓
	days since last purchase		✓
	number of sessions		
	number of devices		
	device sequence score		
	switch probability		

6.5.1 Experimental Setup

In this section, we discuss the feature sets which we use in the experiments for the anonymous and the identified setting, the models on which we test the features, and the evaluation setup.

Feature sets. We start by designing a set of features for purchase prediction in identified and anonymous user settings. Since our initial analysis demonstrated that about 56% of all sessions are anonymous (see Table 6.1), it is worth to pay special attention to this category. Based on the findings obtained thus far and on an analysis of best-performing features available in the literature (Hop, 2013; Lee et al., 2015; Niu et al., 2017; Seippel, 2018), we compile a feature set presented in the Table 6.7.

We categorize features into two classes: *session features* and *customer history features*. We derive session features from the information of the given session and base customer history features on the information from previous sessions of the given customer.

Since we run experiments in the anonymous and the identified setting, we use different feature sets for each setting. In the anonymous setting, the information about the customer is not available and, therefore, we can only use session features. On the other hand, when a customer is identified, we can use both session and customer history features. The feature set contains both static and dynamic features. Dynamic

features can change throughout the session, whereas static features remain constant.

Models. Next, we select models on which we evaluate the features discovered during the dataset analysis. As our primary model, we use a production-ready classifier. This is a random forest (RF) with a baseline feature set as described in Table 6.7. Additionally, to showcase the general utility of our feature set, we experiment on additional models. After reviewing previous work in the domain of purchase prediction (see Section 6.6), we choose the following models for our experiments: logistic regression (LR), K-nearest neighbors (KNN), support vector machines (SVM), neural classifier, and gradient boosted decision tree (GBDT). Each model is trained on the baseline and extended feature set in both settings.

Prediction setup. Since we want to explore how models' performances change across sessions, we select points of a session for which we predict the probability of purchase.

We define a point by the number of pages opened in the session up until the point of prediction. Overall, we select 11 points of measurement. The first point is at the very beginning of the session when the user did not open any pages yet. At this point, the classifier makes a prediction based solely on static features. The following point of measurement is right after the user opened the first page. The subsequent nine points happen after the next nine pages. To make the evaluation possible and to ensure that we do not predict for the very last session page, we filter out sessions with fewer than 12 pages, with 2 pages as a buffer. The buffer is there to avoid the situation when the model predicts at the very end of a session when the outcome is clear. Therefore, we filter out all the sessions which are less than 12 pages long. For example, in step 2 we only have a session with at least 12 actions, which is a hard setting.

Evaluation setup. For both settings, we evaluate model performance with 10-fold cross-validation. To account for class imbalance, we set class weights to be inversely proportional to class frequencies and use F_1 score as a primary evaluation metric.

6.5.2 Prediction for Anonymous Users

First, we evaluate how the added features influence model performance in the anonymous setting, where the user is not known.

Setup. For the anonymous setting, we sampled 22,982 sessions. We use the data to create a feature set for the baseline model and our model. In the anonymous setting, there is no available information about customer history, therefore, we only use session features (see Table 6.7. As can be seen from the table, the baseline feature set comprises four dynamic features, whereas the extended feature set offers five extra features. Since we predict purchase for different points in the session, we compute all dynamic features for a particular session point on which we evaluate. For each session

point, we train the baseline and extended model on the obtained feature sets.

Results. The results in Figure 6.7 show that the additional features boost model performance across all session steps. The performance boost is especially significant at step 0 when a customer has not opened any pages yet. In general, tree-based models (RF and GBDT) and the neural classifier demonstrate the highest scores across all steps. The models are followed by SVM, LR, and KNN classifiers.

The performance of all the models with the baseline feature set improved on step 1. The gain can be explained by the introduction of dynamic session features (step 0 means that the user did not open any pages yet, hence, no dynamic session features). Conversely, for models with the extended feature set, the introduction of dynamic features on step 1 does not significantly increase the performance. After step 1, models' performances reach a plateau.

6.5.3 Prediction of Identified Users

Second, we test models' performances with baseline and extended feature sets in the identified setting, when the user is known.

Setup. For the identified setting, we sampled 6,319 sessions. The feature set for this setting includes session and customer history features (see Table 6.7). During our experiments, we found out that information about the previous session device (such as device type and conversion rate) decrease model performance, so we excluded those features from the training and evaluation sets. This can be explained by the fact that the information about what kind of device users previously used and what was the probability of purchase on that device is not relevant for predicting purchase on the current device. In analogy with the anonymous setting, we prepare feature sets for each of the eleven session points and train and evaluate the models with the baseline and extended feature sets.

Results. Figure 6.7 shows the performance of the models. Overall, the performance of all models for both baseline and extended feature sets and across all steps stays around 96%. The only exception is the k-nearest neighbors classifier where adding extra features on step 0 increases the model's performance by 6.74%. On step 1, however, the gain from the extended feature set is not present. This can be explained by the introduction of the dynamic session features.

6.5.4 Feature Importance Analysis

The experimental results raise a natural question that is 'Which features contribute most to model performance in both settings?' To answer this question, we look at the

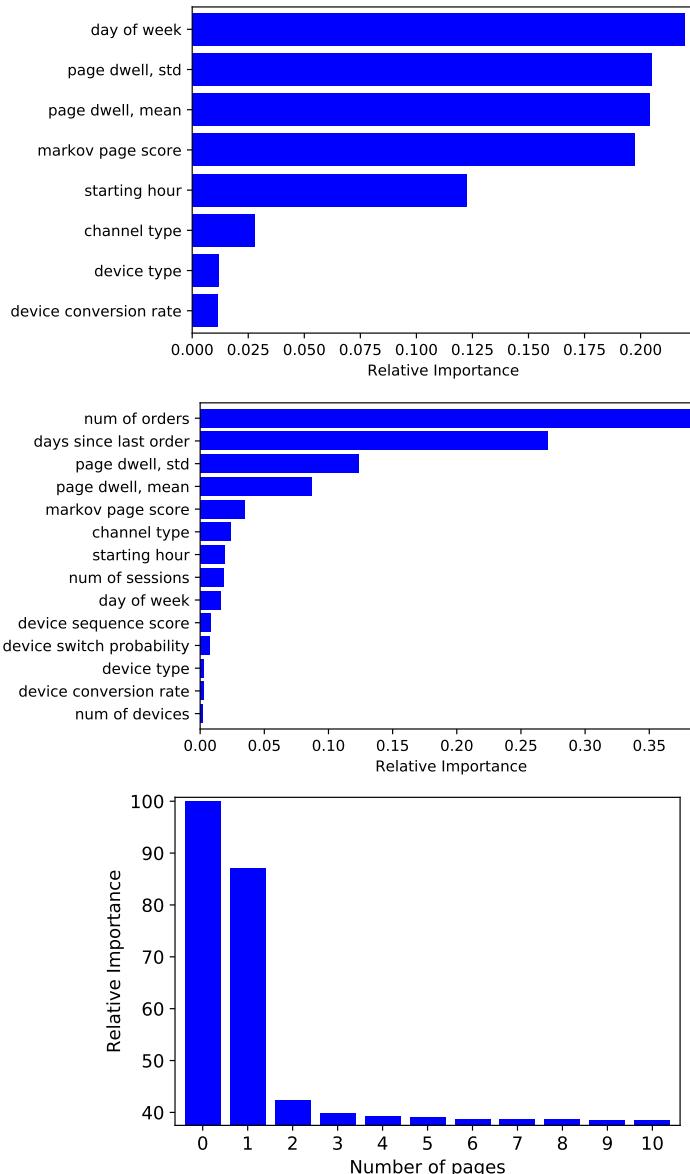


Figure 6.6: Feature importance for the Random Forest in the anonymous setting (top) and the identified setting (center), as well as summed for static features in the anonymous setting (bottom).

feature importance scores of a production-ready classifier which shows one of the best performances in both settings, random forest.

Figure 6.6 (top) demonstrates that in anonymous setting, day of the week is the feature with the highest importance. It is followed by three dynamic features (standard deviation and mean of page dwell time, and Markov page sequence score), and four static features (starting hour, channel type, device type and conversion rate).

Figure 6.6 (center) shows that in the identified user setting, number of previous orders, and number of days since last order are the features with the highest relative importance. Both features describe user historical purchasing behavior what can explain their high relative importance. The features are followed by three dynamic

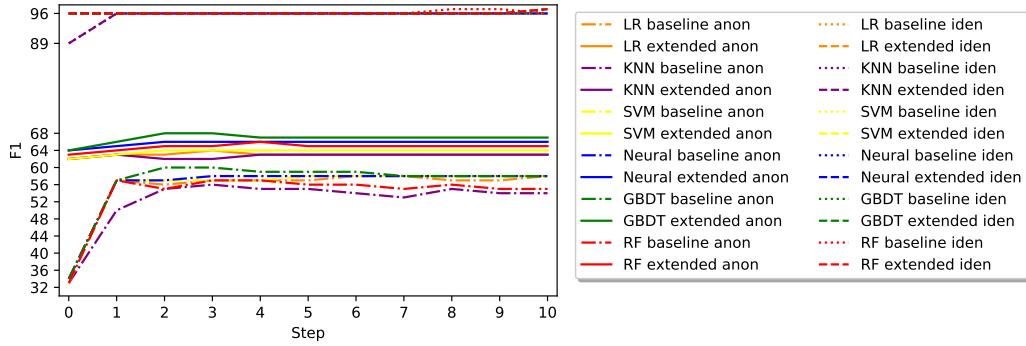


Figure 6.7: Experimental results in anonymous (anon) and identified (iden) setting, across different session steps, F_1 .

features (standard deviation and mean of page dwelling time, and Markov page sequence score), which also have relatively high importance in the anonymous setting. The high relative importance of the dynamic session features (standard deviation and mean of page dwelling time, and Markov page sequence score) in both settings explain the gain all models with baseline feature set got on step 1 in the anonymous setting (see Figure 6.7).

Next, we determine how static feature importance changes across sessions. We consider the importance in the anonymous setting because the introduction of dynamic features in this setting showed an improvement. Figure 6.6 (bottom) shows that static session feature importance decrease as the session evolves, which entails that the importance of dynamic features increases. On step 0 the cumulative importance of static features is 100% because there are no dynamic features introduced. However, from step 1 the relative importance starts to drop. The figure supports the hypothesis that as the session progresses dynamic features become more important.

6.6 RELATED WORK

6.6.1 E-Commerce User Purchase Behavior Analysis

Research on understanding online users' purchasing behavior has been ongoing since the very beginning of e-commerce (Bellman et al., 1999). Studies have investigated user motivation (Bellman et al., 1999), factors that influence e-commerce adoption (O'cass and Fenech, 2003), as well as purchasing behavior (Brown et al., 2003; Hsu et al., 2006), with a focus on perceived security (Salisbury et al., 2001; George, 2002), the decision-making process (Senecal et al., 2005), and purchaser profiles (Swinyard and Smith, 2004; Hernández et al., 2011). Besides, there has also been work on user behavior on content discovery platforms and its relationship to subsequent purchases (Lo et al., 2016) as well as work dedicated to the identification of a taxonomy of product search

intents and the prediction of user satisfaction (Su et al., 2018).

Unlike previous work, our study focuses on the exploration of user purchasing behavior by comparing purchase vs. non-purchase sessions. Besides, we analyze the data from the perspective of device types, and explore aspects such as session length and price of the seen products from the perspective of different devices. On top of that, we also look into the way customer switches between devices.

6.6.2 Purchase Prediction in E-Commerce

The problem of e-commerce user behavior modeling has been studied from various angles, such as building multiple classifiers based on genetic algorithms (Kim et al., 2003), mining purchase patterns with association rules and using those patterns for purchase prediction (Suh et al., 2004). Research has been focused on creating models robust to noise in session data (Agichtein et al., 2006), and using a recurrent neural network to predict customer behaviour (Lang and Rettenmeier, 2017).

Sismeiro and Bucklin (2004) predict purchasing task completion for a given user who completed at least one task earlier, whereas Cheng et al. (2017) explore user behavior on a content discovery platform to determine intent specificity and time in the future when a purchase is estimated to take place. Some work in the field focuses on using queries for purchasing behavior modeling. For instance, Dai et al. (2006) predict purchase based on input query. Besides using general session data, there has been work that incorporates demographic data and perceived attributes (Young Kim and Kim, 2004), scrolling and mouse movements (Guo and Agichtein, 2010), payment data (Wen et al., 2018), log-trace data (Tao et al., 2019), and phone touch actions (Guo et al., 2019a). There has been work on analyzing behavioral patterns and the exploration of different model architectures. In particular, support vector machines, K-nearest neighbor approach, random forest, and logistic regression were used (Lee et al., 2015; Suchacka et al., 2015; Niu et al., 2017).

Unlike previous work in this domain, our study focuses on purchase prediction with two types of users, identified and anonymous. Therefore, we develop two models, run them in two settings, and evaluate their results. The possibility to experiment with identified users also allows us to leverage information from previous user sessions, such as user purchasing history and the number of devices a user owns. In contrast, anonymous users contribute to a higher share of traffic, which makes it important to understand their behavior too. Additionally, we explore how the relevance of dynamic and static features changes as a session progresses.

6.7 DISCUSSION AND CONCLUSION

In this study, we have carried out an analysis of user purchase intent in e-commerce. We have analyzed four weeks of session logs from a European e-commerce platform to identify signals in user behavior that can imply purchase intent. We have considered aspects such as session length, day of the week, and session start hour, as well as information about device, channel, and queries.

In the second part of our study, we have analyzed the relevance of the discovered signals by running a series of experiments aimed at purchase intent prediction in the anonymous and identified settings. We tested the features on random forest, the model which fits production requirements. Additionally, we tested the features on five other models. The experiments demonstrated the value of the features that we engineered based on our insights into the data. We explored which features contribute to performance improvement.

One of the implications of our study is enhanced understanding of purchasing user behavior in e-commerce. Understanding the behavior is the first step towards modeling it, as we demonstrated in the second part of the chapter. Modeling user behavior can contribute towards reducing friction in the customer journey and, therefore, to better customer experience. Besides, we explored the topic of detecting the purchase intent of anonymous users. We showed that, while anonymous users contribute to more than half of the traffic, their user intent is harder to detect because all the predictions have to be made without knowledge about the prior behavior.

Our research has several limitations; one of them is limited generalizability. Even though the data we use in our study comes from a dominant e-commerce platform, it is still only one platform. Hence, it would be interesting to verify the findings against other e-commerce platforms and explore the differences. Moreover, we sampled four weeks of data, thereby introducing a sample bias that could make our findings sensitive to unknown temporal or seasonal patterns. Therefore, it would be interesting to explore if expanding our dataset will lead to new insights. For example, if we had several months of data, we could explore how user purchase intent changes across different months or seasons. Furthermore, we evaluated our purchase intent prediction models in an offline setting. The next logical step is to evaluate them in an online setting.

Future research on the topic includes several directions. First, there is an opportunity to continue research into general purchase behavior analysis and modeling in e-commerce. It would be interesting to explore more aspects of purchasing behavior and try out more models. Another direction for further research concerns predicting purchase intent for anonymous users. Another exciting direction for further research includes modeling device-specific purchase behavior. It can include both relatively

common devices such as PC, smartphone, and tablet, and relatively less popular and studied devices such as TV or game console.

Therefore, our conclusion for RQ5 is that facilitating product retrieval by predicting purchase intent in a cross-device setting involves analyzing behavioral signals from user session logs. Factors such as session duration, timing, device type, and user queries significantly correlate with purchase intent and can be leveraged early in user sessions to anticipate purchasing decisions. Predictive insights are particularly challenging among anonymous users due to the lack of prior behavioral data. Device transitions play a crucial role in predicting purchase behavior, underscoring the necessity of cross-device analysis in e-commerce platforms.

7

EXTENDING CLIP FOR CATEGORY-TO-IMAGE RETRIEVAL

In this chapter we focus on the problem of alignment of textual descriptions of categories with corresponding visual representations for categories of varying granularity. In many cases in such scenarios, textual representations of categories may not adequately capture the visual nuances of associated images, leading to mismatches and suboptimal retrieval results. Hence, ability of information retrieval (IR) systems to bridge the semantic gap across modalities in such scenario facilitates a more intuitive and efficient user experience. Motivated by this problem, we propose a task of category-to-image (CTI). The task involves retrieving a ranked list of relevant images corresponding to a given category sampled from a category tree. Hence, we address the following research question:

RQ6: How do multimodal document representation, encompassing text, image, and attribute data, impact the performance on the category-to-image retrieval in the context of categories of varying granularity?

To answer this research question we formulate the CTI retrieval task and prepare a dataset containing textual descriptions, images, and attribute information of products across categories of varying granularity. We design a multimodal retrieval model that integrates information from text, image, and attribute modalities, and compare its performance against baseline models in experimental settings. Our findings help us understand how combining various modalities impact models performance on the task.

This chapter was published at the 44th European Conference on Information Retrieval (ECIR 2022) under the title “Extending CLIP for Category-to-image Retrieval in E-commerce” (Hendriksen et al., 2022).

7.1 INTRODUCTION

Multimodal retrieval is an important, understudied problem in e-commerce (Tsagkias et al., 2020). Even though e-commerce products are associated with rich multi-modal information, research currently focuses mainly on textual and behavioral signals to support product search and recommendation. The majority of prior work in multimodal retrieval for e-commerce focuses on applications in the fashion domain, such as recommendation of fashion items (Lin et al., 2019) and cross-modal fashion retrieval (Laenen and Moens, 2019; Goei et al., 2021). In the more general e-commerce domain, multimodal retrieval has not been explored that well yet (Hewawalpita and Perera, 2019; Li et al., 2020b). The multimodal problem on which we focus is motivated by the importance of category information in e-commerce. Product category trees are a key component of modern e-commerce as they assist customers when navigating across large and dynamic product catalogues (Wirojwatanakul and Wangperawong, 2019; Tagliabue et al., 2020; Kondylidis et al., 2021). Yet, the ability to retrieve an image for a given product category remains a challenging task mainly due to noisy category and product data, and the size and dynamic character of product catalogues (Laenen et al., 2018; Tsagkias et al., 2020).

The category-to-image retrieval task. We introduce the problem of retrieving a ranked list of relevant images of products that belong to a given category, which we call the *category-to-image* (CTI) retrieval task. Unlike image classification tasks that operate on a predefined set of classes, in the CTI retrieval task we want to be able not only to understand which images belong to a given category but also to generalize towards unseen categories. Consider the category “Home decor.” A CTI retrieval should output a ranked list of k images retrieved from the collection of images that are relevant to the category, which could be anything from images of carpets to an image of a clock or an arrangement of decorative vases. Use cases that motivate the CTI retrieval task include (i) the need to showcase different categories in search and recommendation results (Tsagkias et al., 2020; Tagliabue et al., 2020; Kondylidis et al., 2021); (ii) the task can be used to infer product categories in the cases when product categorical data is unavailable, noisy, or incomplete (Yashima et al., 2016); and (iii) the design of cross-categorical promotions and product category landing pages (Nielsen et al., 2000).

The CTI retrieval task has several key characteristics: (i) we operate with categories from non-fixed e-commerce category trees, which range from very general (such as “Automotive” or “Home & Kitchen”) to very specific ones (such as “Helmet Liners” or “Dehumidifiers”). The category tree is not fixed, therefore, we should be able to generalize towards unseen categories; and (ii) product information is highly multimodal in nature; apart from category data, products may come with textual, visual, and attribute information.

A model for CTI retrieval. To address the CTI retrieval task, we propose a model that leverages image, text, and attribute information, CLIP-ITA. CLIP-ITA extends upon Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021). CLIP-ITA extends CLIP with the ability to represent attribute information. Hence, CLIP-ITA is able to use textual, visual, and attribute information for product representation. We compare the performance of CLIP-ITA with several baselines such as unimodal BM25, bimodal zero-shot CLIP, and MPNet (Song et al., 2020). For our experiments, we use the XMarket dataset that contains textual, visual, and attribute information of e-commerce products (Bonab et al., 2021).

Research questions and contributions. We address the following research questions: **(RQ6.1)** How do baseline models perform on the CTI retrieval task? Specifically, how do unimodal and bi-modal baseline models perform? How does the performance differ w.r.t. category granularity? **(RQ6.2)** How does a model, named CLIP-I, that uses product image information for building product representations impact the performance on the CTI retrieval task? **(RQ6.3)** How does CLIP-IA, which extends CLIP-I with product attribute information, perform on the CTI retrieval task? **(RQ6.4)** And finally, how does CLIP-ITA, which extends CLIP-IA with product text information, perform on the CTI task?

Our main contributions are: (i) We introduce the novel task of CTI retrieval and motivate it in terms of e-commerce applications. (ii) We propose CLIP-ITA, the first model specifically designed for this task. CLIP-ITA leverages multimodal product data such as textual, visual, and attribute data. On average, CLIP-ITA outperforms CLIP-I on all categories by 217% and CLIP-IA by 269%. We share our code and experimental settings to facilitate reproducibility of our results.

7.2 RELATED WORK

7.2.1 Learning Multimodal Embeddings.

Contrastive pre-training has been shown to be highly effective in learning joined embeddings across modalities (Radford et al., 2021). By predicting the correct pairing of image-text tuples in a batch, the CLIP model can learn strong text and image encoders that project to joint space. This approach to learning multimodal embeddings offers key advantages over approaches that use manually assigned labels as supervision: (i) the training data can be collected without manual annotation; real-world data in which image-text pairs occur can be used; (ii) models trained in this manner learn more general representations that allow for zero-shot prediction. These advantages are appealing for e-commerce, as most public multimodal e-commerce datasets pri-

marily focus on fashion only (Bonab et al., 2021); being able to train from real-world data avoids the need for costly data annotation.

We build on CLIP by extending it to category-product pairs, taking advantage of its ability to perform zero-shot retrieval for a variety of semantic concepts.

7.2.2 Multimodal Image Retrieval

Early work in image retrieval grouped images into a restricted set of semantic categories and allowed users to retrieve images by using category labels as queries (Smeulders et al., 2000). Later work allowed for a wider variety of queries ranging from natural language (Hu et al., 2016; Vo et al., 2019), to attributes (Nagarajan and Grauman, 2018), to combinations of multiple modalities such as title, description, and tags (Thomee et al., 2016). Across these multimodal image retrieval approaches we find three common components: (1) an image encoder, (2) a query encoder, and (3) a similarity function to match the query to images (Radford et al., 2021; Gupta et al., 2020). Depending on the focus of the work some components might be pre-trained, whereas the others are optimized for a specific task.

In our work, we rely on pre-trained image and text encoders but learn a new multimodal composite of the query to perform CTI retrieval.

7.2.3 Multimodal Retrieval in E-Commerce

Prior work on multimodal retrieval in e-commerce has been mainly focused on cross-modal retrieval for fashion (Zoghbi et al., 2016; Laenen et al., 2017; Goei et al., 2021). Other related examples include outfit recommendation (Lin et al., 2019; Laenen and Moens, 2020; Li et al., 2020c) Some prior work on interpretability for fashion product retrieval proposes to leverage multimodal signals to improve explainability of latent features (Liao et al., 2018; Yang et al., 2019). Tautkute et al. (2019) propose a multimodal search engine for fashion items and furniture. When it comes to combining signals for improving product retrieval, Yim et al. (2018) propose to combine product images, titles, categories, and descriptions to improve product search, Yamaura et al. (2019) propose an algorithm that leverages multimodal product information for predicting a resale price of a second-hand product.

Unlike prior work on multimodal retrieval in e-commerce that mainly focuses on fashion data, we focus on creating multimodal product representations for the general e-commerce domain.

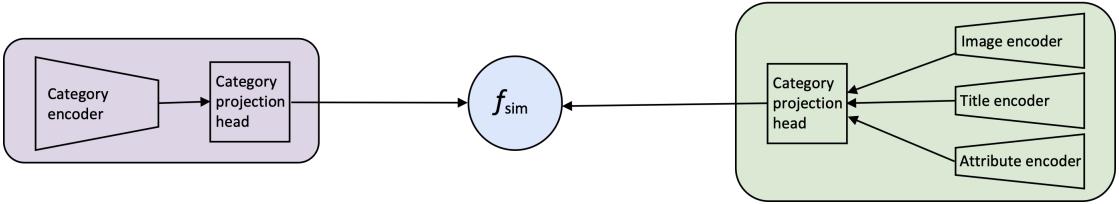


Figure 7.1: Overview of CLIP-ITA. The category encoding pipeline is in purple; the category information pipeline in green; f_{sim} is a cosine similarity function.

7.3 APPROACH

7.3.1 Task Definition

We follow the same notation as in (Zhang et al., 2022c). The input dataset can be presented as category-product pairs $(\mathbf{x}_c, \mathbf{x}_p)$, where \mathbf{x}_c represents a product category, and \mathbf{x}_p represents information about a product that belongs to the category \mathbf{x}_c . The product category \mathbf{x}_c is taken from the category tree T and is represented as a category name. The product information comprises titles \mathbf{x}_t , images \mathbf{x}_i , and attributes \mathbf{x}_a , i.e., $\mathbf{x}_p = \{\mathbf{x}_i, \mathbf{x}_t, \mathbf{x}_a\}$.

For the CTI retrieval task, we use the target category name \mathbf{x}_c as a query and we aim to return a ranked list of top- k images that belong to the category \mathbf{x}_c .

7.3.2 CLIP-ITA

Figure 7.1 provides a high-level view of CLIP-ITA. CLIP-ITA projects category \mathbf{x}_c and product information \mathbf{x}_p into a d -dimensional multimodal space where the resulting vectors are respectively \mathbf{c} and \mathbf{p} . The category and product information is processed by a category encoding pipeline and product information encoding pipeline. The core components of CLIP-ITA are the encoding and projection modules. The model consists out of four encoders: a category encoder, an image encoder, a title encoder, and an attribute encoder. Besides, CLIP-ITA comprises two non-linear projection heads: the category projection head and the multimodal projection head.

While several components of CLIP-ITA are based on CLIP (Radford et al., 2021), CLIP-ITA differs from CLIP in three important ways: (i) unlike CLIP, which operates on two encoders (textual and visual), CLIP-ITA extends CLIP towards a category encoder, image encoder, textual encoder, and attribute encoder; (ii) CLIP-ITA features two projection heads, one for the category encoding pipeline, and one for the product information encoding pipeline; and (iii) while CLIP is trained on text-image pairs, CLIP-ITA is trained on category-product pairs, where product representation is multimodal.

Category Encoding Pipeline. The *category encoder* (f_c) takes as input category name

\mathbf{x}_c and returns its representation \mathbf{h}_c . More specifically, we pass the category name \mathbf{x}_c through the category encoder f_c :

$$\mathbf{h}_c = f_c(\mathbf{x}_c). \quad (7.1)$$

To obtain this representation, we use a pre-trained MPNet model (Song et al., 2020). After passing category information through the category encoder, we feed it to the category projection head. The *category projection head* (g_c) takes as input a query representation \mathbf{h}_c and projects it into d -dimensional multi-modal space:

$$\mathbf{c} = g_c(\mathbf{h}_c), \quad (7.2)$$

where $\mathbf{c} \in \mathbb{R}^d$.

Product Encoding Pipeline. The product information encoding pipeline represents three encoders, one for every modality, and a product projection head. The *image encoder* (f_i) takes as input a product image \mathbf{x}_i aligned with the category \mathbf{x}_c . Similarly to the category processing pipeline, we pass the product image \mathbf{x}_i through the image encoder:

$$\mathbf{h}_i = f_i(\mathbf{x}_i). \quad (7.3)$$

To obtain the image representation \mathbf{h}_i , we use a pre-trained Vision Transformer from CLIP model. The *title encoder* (f_t) takes a product title \mathbf{x}_t as input and returns a title representation \mathbf{h}_t :

$$\mathbf{h}_t = f_t(\mathbf{x}_t). \quad (7.4)$$

Similarly to the category encoder f_c , we use pre-trained MPNet to obtain the title representation \mathbf{h}_t . The *attribute encoder* (f_a) is a network that takes as input a set of attributes $\mathbf{x}_a = \{a_1, a_2, \dots, a_n\}$ and returns their joint representation:

$$\mathbf{h}_a = f_a(\mathbf{x}_a) = \frac{1}{n} \sum_{i=1}^n f_a(\mathbf{x}_{ai}). \quad (7.5)$$

Similarly to the category encoder f_c and title encoder f_t , we obtain representation of each attribute with the pre-trained MPNet model. After obtaining title, image and attribute representations, we pass the representations into the product projection head. The *product projection head* (g_p) takes as input a concatenation of the image representation \mathbf{h}_i , title representation \mathbf{h}_t , and attribute representation \mathbf{h}_a and projects the resulting vector $\mathbf{h}_p = concat(\mathbf{h}_i, \mathbf{h}_t, \mathbf{h}_a)$ into multimodal space:

$$\mathbf{p} = g_p(\mathbf{h}_p) = g_p(concat(\mathbf{h}_i, \mathbf{h}_t, \mathbf{h}_a)), \quad (7.6)$$

where $\mathbf{p} \in \mathbb{R}^d$.

Loss Function. We train CLIP-ITA using bidirectional contrastive loss (Zhang et al., 2022c). The loss is a weighted combination of two losses: a category-to-product

contrastive loss and a product-to-category contrastive loss. In both cases the loss is the InfoNCE loss (Oord et al., 2018). Unlike prior work that focuses on a contrastive loss between inputs of the same modality (He et al., 2020; Chen et al., 2020a) and on corresponding inputs of two modalities (Zhang et al., 2022c), we use the loss to work with inputs from textual modality (category representation) vs. a combination of multiple modalities (product representation). We train CLIP-ITA on batches of category-product pairs $(\mathbf{x}_c, \mathbf{x}_p)$ with batch size β . For the j -th pair in the batch, the category-to-product contrastive loss is computed as follows:

$$\ell_j^{(c \rightarrow p)} = -\log \frac{\exp(f_{sim}(\mathbf{c}_j, \mathbf{p}_j)/\tau)}{\sum_{k=1}^{\beta} \exp(f_{sim}(\mathbf{c}_j, \mathbf{p}_k)/\tau)}, \quad (7.7)$$

where $f_{sim}(\mathbf{c}_i, \mathbf{p}_i)$ is the cosine similarity, and $\tau \in \mathbb{R}^+$ is a temperature parameter. Similarly, the product-to-category loss is computed as follows:

$$\ell_j^{(p \rightarrow c)} = -\log \frac{\exp(f_{sim}(\mathbf{p}_j, \mathbf{c}_j)/\tau)}{\sum_{k=1}^{\beta} \exp(f_{sim}(\mathbf{p}_j, \mathbf{c}_k)/\tau)}. \quad (7.8)$$

The resulting contrastive loss is a combination of the two above-mentioned losses:

$$\mathcal{L} = \frac{1}{\beta} \sum_{j=1}^{\beta} (\lambda \ell_j^{(p \rightarrow c)} + (1 - \lambda) \ell_j^{(c \rightarrow p)}), \quad (7.9)$$

where β represents the batch size and $\lambda \in [0, 1]$ is a scalar weight.

7.4 EXPERIMENTAL SETUP

7.4.1 Dataset.

We use the XMarket dataset recently introduced by Bonab et al. (2021) that contains textual, visual, and attribute information of e-commerce products as well as a category tree. For our experiments, we select 38,921 products from the US market. Category information is represented as a category tree and comprises 5,471 unique categories across nine levels. Level one is the most general category level, level nine is the most specific level. Every product belongs to a subtree of categories $t \in T$. In every subtree t , each parent category has only one associated child category. The average subtree depth is 4.63 (minimum: 2, maximum: 9). Because every product belongs to a subtree of categories, the dataset contains 180,094 product-category pairs in total. We use product titles as textual information and one image per product as visual information. The attribute information comprises 228,368 attributes, with 157,049 unique. On average, every product has 5.87 attributes (minimum: 1, maximum: 24).

7.4.2 Evaluation Method

To investigate how model performance changes w.r.t. category granularity, for every product in the dataset, \mathbf{x}_p , and the corresponding subtree of categories to which the product belongs, t , we train and evaluate the model performance in three settings: (i) *all categories*, where we randomly select one category from the subtree t ; (ii) *most general category*, where we use only the most general category of the subtree t , i.e., the root; and (iii) *most specific category*, where we use the most specific category of the subtree t . In total, there are 5,471 categories in all categories setup, 34 categories in the most general category, and 4,100 in the most specific category setup. We evaluate every model on category-product pairs $(\mathbf{x}_c, \mathbf{x}_p)$ from the test set. We encode each category and a candidate product data by passing them through category encoding and product information encoding pipelines. For every category \mathbf{x}_c we retrieve the top- k candidates ranked by cosine similarity w.r.t. the target category \mathbf{x}_c .

Metrics. To evaluate model performance, we use Precision@K where $K = \{1, 5, 10\}$, mAP@K where $K = \{5, 10\}$, and R-precision.

Baselines. Following (Wang et al., 2021c; Shen et al., 2021; Dai et al., 2020) we use BM25, MPNet, CLIP as our baselines.

Four experiments. We run four experiments, corresponding to our research questions as listed at the end of Section 7.1. In *Experiment 1* we evaluate the baselines on the CTI retrieval task (RQ6.1). We feed BM25 corpora that contain textual product information, i.e., product titles. We use MPNet in a zero-shot manner. For all the products in the dataset, we pass the product title \mathbf{x}_t through the model. During the evaluation, we pass a category \mathbf{x}_c expressed as textual query through MPNet and retrieve top- k candidates ranked by cosine similarity w.r.t. the target category \mathbf{x}_c . We compare categories of the top- k retrieved candidates with the target category \mathbf{x}_c . Besides, we use pre-trained CLIP in a zero-shot manner with a Text Transformer and a Vision Transformer (ViT) (Dosovitskiy et al., 2021) as configuration. We pass the product images \mathbf{x}_i through the image encoder. For evaluation, we pass a category \mathbf{x}_c through the text encoder and retrieve top- k image candidates ranked by cosine similarity w.r.t. the target category \mathbf{x}_c . We compare categories of the top- k retrieved images with the target category \mathbf{x}_c .

In *Experiment 2* we evaluate image-based product representations (RQ6.2). After obtaining results with CLIP in a zero-shot setting, we build product representations by training on e-commerce data. First, we investigate how using product image data for building product representations impacts performance on the CTI retrieval task. To introduce visual information, we extend CLIP in two ways: (1) We use ViT from CLIP as image encoder f_i . We add product projection head g_p that takes as an input product visual information $\mathbf{x}_i \in \mathbf{x}_p$. (2) We use the text encoder from MPNet as cate-

gory encoder f_c ; we add a category projection head g_c on top of category encoder f_c thereby completing category encoding pipeline (see Figure 7.1). We name the resulting model CLIP-I. We train CLIP-I on category-product pairs $(\mathbf{x}_c, \mathbf{x}_p)$ from the training set. Note that $\mathbf{x}_p = \{\mathbf{x}_i\}$, i.e., we only use visual information for building product representations.

In *Experiment 3*, we evaluate image- and attribute-based product representations (RQ6.3). We extend CLIP-I by introducing attribute information to the product information encoding pipeline. We add an attribute encoder f_a through which we obtain a representation of product attributes, \mathbf{h}_a . We concatenate the resulting attribute representation with image representation $\mathbf{h}_p = concat(\mathbf{h}_i, \mathbf{h}_a)$ and pass the resulting vector to the product projection head g_p . Thus, the resulting product representation \mathbf{p} is based on both visual and attribute product information. We name the resulting model CLIP-IA. We train CLIP-IA on category-product pairs $(\mathbf{x}_c, \mathbf{x}_p)$ where $\mathbf{x}_p = \{\mathbf{x}_i, \mathbf{x}_a\}$, i.e., we use visual and attribute information for building product representation.

In *Experiment 4*, we evaluate image-, attribute-, and title-based product representations (RQ6.4). We investigate how extending the product information processing pipeline with the textual modality impacts performance on the CTI retrieval task. We add title encoder f_t to the product information processing pipeline and use it to obtain title representation \mathbf{h}_t . We concatenate the resulting representation with product image and attribute representations $\mathbf{h}_p = concat(\mathbf{h}_i, \mathbf{h}_t, \mathbf{h}_a)$. We pass the resulting vector to the product projection head g_p . The resulting model is CLIP-ITA. We train and test CLIP-ITA on category-product pairs $(\mathbf{x}_c, \mathbf{x}_p)$ where $\mathbf{x}_p = \{\mathbf{x}_i, \mathbf{x}_a, \mathbf{x}_t\}$, i.e., we use visual, attribute, and textual information for building product representations.

Implementation details. We train every model for 30 epochs, with a batch size $\beta = 8$ for most general categories, $\beta = 128$ — for most specific categories and all categories. For loss function, we set $\tau = 1$, $\lambda = 0.5$. We implement every projection head as non-linear MLPs with two hidden layers, GELU non-linearities (Hendrycks and Gimpel, 2016) and layer normalization (Ba et al., 2016). We optimize both heads with the AdamW optimizer (Loshchilov and Hutter, 2019).

7.5 EXPERIMENTAL RESULTS

7.5.1 Baselines.

Following RQ6.1, we start by investigating how do baselines perform on cross-modal retrieval. Besides, we investigate how does the performance on the task differs between the unimodal and the bimodal approach.

The results are shown in Table 7.1. When evaluating on all categories, all the base-

Table 7.1: Results of Experiments 1–4. The best performance is highlighted in bold.

Model	P@1	P@5	P@10	MAP@5	MAP@10	R-precision
All categories (5,471)						
BM25 (Jones et al., 2000)	0.01	0.01	0.01	0.01	0.01	0.01
CLIP (Radford et al., 2021)	0.01	0.02	0.02	0.03	0.04	0.02
MPNet (Song et al., 2020)	0.01	0.06	0.06	0.07	0.09	0.05
CLIP-I (Ours)	3.3	3.8	3.79	6.81	7.25	3.67
CLIP-IA (Ours)	2.5	3.34	3.29	5.95	6.24	3.27
CLIP-ITA (Ours)	9.9	13.27	13.43	20.3	20.53	13.42
Most general category (34)						
BM25 (Jones et al., 2000)	2.94	4.71	4.71	8.33	8.28	4.48
CLIP (Radford et al., 2021)	11.76	12.35	11.76	16.12	15.18	9.47
MPNet (Song et al., 2020)	14.70	15.8	15.01	18.44	18.78	9.35
CLIP-I (Ours)	17.85	17.14	16.78	19.88	20.14	13.02
CLIP-IA (Ours)	21.42	21.91	22.78	25.59	26.29	20.74
CLIP-ITA (Ours)	35.71	30.95	30.95	35.51	34.28	25.79
Most specific category (4,100)						
BM25 (Jones et al., 2000)	0.02	0.02	0.01	0.01	0.01	0.01
CLIP (Radford et al., 2021)	11.92	9.81	9.23	15.12	14.95	8.14
MPNet (Song et al., 2020)	33.36	28.56	26.93	37.43	36.77	25.29
CLIP-I (Ours)	14.06	12.11	11.53	18.24	17.9	11.22
CLIP-IA (Ours)	35.3	30.21	29.32	39.93	39.27	28.86
CLIP-ITA (Ours)	45.85	41.04	40.02	50.04	49.87	39.69

lines perform poorly. For the most general category setting, MPNet outperforms CLIP on all metrics except R-precision. The most prominent gain is for Precision@10 where MPNet outperforms CLIP by 28%. CLIP outperforms BM25 on all metrics. For the most specific category setting, MPNet performance is the highest, BM25 — the lowest. In particular, MPNet outperforms CLIP by 211% in Precision@10. Overall, MPNet outperforms CLIP and both models significantly outperform BM25 for both most general and most specific categories. However, when evaluation is done on all categories, the performance of all models is comparable. As an answer to RQ6.1, the results suggest that using information from multiple modalities is beneficial for performance on the task.

7.5.2 *Image-Based Representations*

To address RQ6.2, we compare the performance of CLIP-I with CLIP and MPNet, the best-performing baseline. Table 7.1, shows the experimental results for Experiment 2. The biggest performance gains are obtained in “all categories” setting. However, there, the performance of the baselines was very poor. For the most general categories, CLIP-I outperforms both CLIP and MPNet. For CLIP-I vs. CLIP, we observe the biggest increase of 51% for Precision@1, for CLIP-I vs. MPNet — 39% in R-precision. In the case of the most specific categories, CLIP-I outperforms CLIP but loses to MPNet. Overall, CLIP-I outperforms CLIP in all three settings and outperforms MPNet except the most specific categories. Therefore, we answer RQ6.2 as follows: the results suggest that extension of CLIP by the introduction of product image data for building product representations has a positive impact on performance on cross-modal retrieval.

7.5.3 *Image- and Attribute-Based Representations*

To answer RQ6.3, we compare the performance of CLIP-IA with CLIP-I and the baselines. The results are shown in Table 7.1. When evaluated on all categories, CLIP-IA performs worse than CLIP-I but outperforms MPNet. In particular, CLIP-I obtains the biggest gain relative of 32% on Precision@1 and the lowest gain of 12% on R-precision. For the most general category, CLIP-IA outperforms CLIP-I and MPNet on all metrics. More specifically, we observe the biggest gain of 122% on R-precision over MPNet and the biggest gain of 59% on R-precision for CLIP-I. Similarly, for the most specific category, CLIP-IA outperforms both CLIP-I and MPNet. We observe the biggest relative gain of 138% over CLIP-I. The results suggest that further extension of CLIP by the introduction of the product image and attribute data for building product representations has a positive impact on performance on cross-modal retrieval, especially when evaluated on most specific categories. Therefore, we answer RQ6.3 positively.

7.5.4 *Image-, Attribute-, and Title-Based Representations*

We compare CLIP-ITA with both CLIP-IA, CLIP-I, and the baselines. The results are shown in Table 7.1. In general, CLIP-ITA outperforms CLIP-I and CLIP-IA and the baselines in all settings. When evaluated on all categories, the maximum relative increase of CLIP-ITA over CLIP-I is 265% in R-precision, the minimum relative increase is 183% in mAP@10. The biggest relative increase of CLIP-ITA performance over CLIP-IA is 310% in Precision@1, the smallest relative increase is 229% in mAP@10. For the most general categories, CLIP-ITA outperforms CLIP-I by 82% and CLIP-IA by 38%.

Table 7.2: Erroneous CLIP-ITA prediction counts for “same tree” vs. “different tree” predictions per evaluation type.

	Same tree	Different tree
All categories	1,655	639
The most general category	2	21
The most specific category	127	1,011
Total	1,786	1,671

For most specific categories, we observe the biggest increase of CLIP-ITA over CLIP-I of 254% in R-precision and the smallest relative increase of 172% on mAP@5. At the same time, the biggest relative increase of CLIP-ITA over CLIP-IA is a 38% increase in R-precision and the smallest relative increase is a 27% increase in mAP@5. Overall, CLIP-ITA wins in all three settings. Hence, we answer RQ6.4 positively.

7.6 ERROR ANALYSIS

7.6.1 Distance between Predicted and Target Categories.

We examine the performance of CLIP-ITA by looking at the pairs of the ground-truth and predicted categories (c, c_p) in cases when the model failed to predict the correct category, i.e., $c \neq c_p$. This allows us to quantify how far off the incorrect predictions lie w.r.t. the category tree hierarchy. First, we examine in how many cases target category c and predicted category c_p belong to the same most general category, i.e., belong to the same category tree; see Table 7.2. In the case of most general categories, the majority of incorrectly predicted categories belong to a tree different from the target category tree. For the most specific categories, about 11% of predicted categories belong to the category tree of the target category. However, when evaluation is done on all categories, 72% of incorrectly predicted cases belong to the same tree as a target category.

Next, we turn to the category-predicted category pairs (c, c_p) where the incorrectly predicted category c_p belongs to the same tree as target category c . We compute the distance d between a category used as a query c and a predicted category c_p . We compute the distance between target category c and a top-1 predicted category c_p as the difference between their respective depths $d(c, c_p) = \text{depth}(c_p) - \text{depth}(c)$. The distance d is positive if the depth of the predicted category is bigger than the depth of the target category, $\text{depth}(c_p) > \text{depth}(c)$, i.e., the predicted category is more specific than the target category. The setup is mirrored for negative distances.

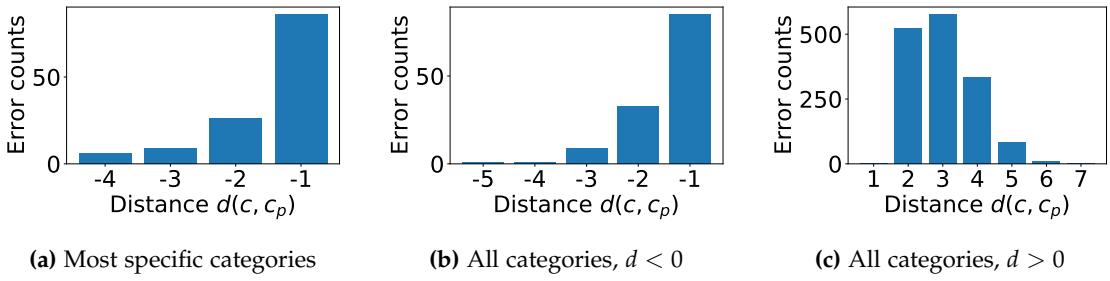


Figure 7.2: Error analysis for CLIP-ITA. Distance between target category c and a predicted category c_p when c and c_p are in the same tree.

See Figure 7.2. We do not plot the results for the most general category because for this setting there are only two cases when target category c and a predicted category c_p were in the same tree. In both cases, predicted category c_p was more general than target category c with distance $d(c, p_c) = 2$. In cases when target category c was sampled from the most specific categories, the wrongly predicted category c_p belonging to the same tree was always more specific than the target category c with the maximum absolute distance between c and c_p , $|d(c, c_p)| = 4$. In 68% of the cases the predicted category was one level above the target category, for 21% $d(c, c_p) = -2$, for 7% $d(c, c_p) = -3$, and for 5% $d(c, c_p) = -4$. For the setting with all categories, in 92% of the cases, the predicted category c_p was more specific than the target category c ; for 8% the predicted category was more general.

Overall, for the most general category and the most specific category, the majority of incorrectly predicted categories are located in a category tree different from the one where the target category was located. For the “all categories” setting, it is the other way around. When it comes to the cases when incorrectly predicted categories are in the same tree as a target category, the majority of incorrect predictions are 1 level more general when the target category is sampled from the most specific categories. For the “all categories” setting, the majority of incorrect predictions belonging to the same tree as the target category were more specific than the target category. Our analysis suggests that efforts to improve the performance of CLIP-ITA should focus on minimizing the (tree-based) distance between the target and predicted category in a category tree. This could be incorporated as a suitable extension of the loss function.

7.6.2 Performance on Seen vs. Unseen Categories

Next, we investigate how well CLIP-ITA generalizes to unseen categories. We split the evaluation results into two groups based on whether the category used as a query was seen during training or not; see Table 7.3. For the most general categories, CLIP-ITA is unable to correctly retrieve an image of the product of the category that was not seen during training at all. For the most specific categories, CLIP-ITA performs

Table 7.3: CLIP-ITA performance on seen vs. unseen categories.

Model	P@1	P@5	P@10	mAP@5	mAP@10	R-precision
All categories (5,471)						
CLIP-ITA (unseen cat.)	13.3	18.56	15.55	19.7	19.65	18.52
CLIP-ITA (seen cat.)	10.48	13.95	14.08	21.65	21.65	14.07
Most general category (34)						
CLIP-ITA (unseen cat.)	0.0	0.0	0.0	0.0	0.0	0.0
CLIP-ITA (seen cat.)	19.23	20.01	17.31	20.41	20.01	15.73
Most specific category (4,100)						
CLIP-ITA (unseen cat.)	27.27	26.44	26.44	27.92	27.92	26.45
CLIP-ITA (seen cat.)	47.83	43.09	42.14	52.41	51.89	41.58

better on seen categories than on unseen categories. We observe the biggest relative performance increase of 85% in mAP@10 and the smallest relative increase of 57% in R-precision. When evaluating on all categories, CLIP-ITA performs on unseen categories better when evaluated on Precision@k (27% higher in Precision@1, 33% higher in Precision@5, 10% increase in Precision@10) and R-precision (relative increase of 32%). Performance on seen categories is better in terms of mAP@k (10% increase for both mAP@5 and mAP@10).

Overall, for the most general and most specific categories, the model performs much better on categories seen during training. For “all categories” setting, however, CLIP-ITA’s performance on unseen categories is better.

7.7 CONCLUSION

We introduced the task of category-to-image retrieval and motivated its importance in the e-commerce scenario. In the CTI retrieval task, we aim to retrieve an image of a product that belongs to the target category. We proposed a model specifically designed for this task, CLIP-ITA. CLIP-ITA extends CLIP, one of the best performing text-image retrieval models. CLIP-ITA leverages multimodal product data such as textual, visual, and attribute data to build product representations. In our experiments, we contrasted and evaluated different combinations of signals from modalities, using three settings: on all categories, the most general, and the most specific categories.

We found that combining information from multiple modalities to build product representation produces the best results on the cross-modal retrieval. CLIP-ITA gives the best performance both on all categories and on the most specific categories. On

the most general categories, CLIP-I, a model where product representation is based on image only, works slightly better. CLIP-I performs worse on the most specific categories and across all categories. For identification of the most general categories, visual information is more relevant. Besides, CLIP-ITA is able to generalize to unseen categories except in the case of most general categories. However, the performance on unseen categories is lower than the performance on seen categories. Even though our work is focused on the e-commerce domain, the findings can be useful for other areas, e.g., digital humanities.

Limitations of our work are due to type of data in the e-commerce domain. In e-commerce, there is typically one object per image and the background is homogeneous, textual information is lengthly and noisy; in the general domain, there is typically more than one object per image, image captions are more informative and shorter. Future work directions can focus on improving the model architecture. It would be interesting to incorporate attention mechanisms into the attribute encoder and explore how it influences performance. Another interesting direction for future work is to evaluate CLIP-ITA on other datasets outside of the e-commerce domain. Future work can also focus on minimizing the distance between the target and predicted category in the category tree.

To summarize, we answer RQ6 by saying that multimodal document representation, integrating text, image, and attribute data, improves CTI retrieval performance across categories of varying granularity. The CLIP-ITA model, designed for this task, demonstrates superior performance by combining multiple modalities. For the most general categories, models using image information alone perform slightly better, highlighting the importance of visual cues. However, for more specific categories, the inclusion of textual and attribute data becomes important, indicating that detailed information is necessary for accurate retrieval as category specificity increases.

REPRODUCIBILITY

To ensure the reproducibility of the findings presented in this chapter, we have made our code publicly accessible at https://github.com/mariyahendriksen/ecir2022_category_to_image_retrieval.

8

CONCLUSION

In this thesis, we have focused on multimodal machine learning for information retrieval from a vision and language perspective. We covered a variety of challenges and proposed novel architectures to address them in the context of dense and sparse retrieval, representation learning and evaluation, and product retrieval.

Dense and Sparse Retrieval In Chapter 2, we focused on assessing the reproducibility of cross-modal retrieval (CMR) results on scene-centric and object-centric datasets. We discovered challenges in replicating CMR performance due to variations in data preprocessing and experimental setups. These findings underscore the need for standardized benchmarks and methodologies to improve consistency in CMR research. Meanwhile, in Chapter 3, we introduced a method for transforming dense vision-language (VL) representations into sparse ones, aiming to enhance computational efficiency without compromising effectiveness. The results highlighted a promising trade-off where sparse models achieved competitive performance with reduced computational demands, demonstrating potential scalability benefits for real-world applications of VL models.

Representation Learning and Evaluation In Chapter 4, we investigated the problem of shortcut learning for VL contrastive learning (CL) representation learning with multiple captions per image and proposed a framework to study the problem in a controlled way. We discovered that contrastive losses tend to prioritize learning easily detectable features shared between the image and all captions, neglecting other relevant information that might be unique to specific captions. In Chapter 5, we shifted our focus to evaluating the brittleness of existing benchmarks when evaluated on the image-text retrieval (ITR) task. We emphasized the limitations of the current evaluation pipeline, advocating for more refined evaluation frameworks. These insights underline the importance of developing comprehensive evaluation protocols that better reflect real-world complexities.

Product Retrieval In Chapter 6, we focused on facilitating product retrieval by predicting user purchase intent through behavioral analysis of e-commerce session logs. The findings identified key behavioral signals that correlate with purchasing decisions. These insights offer practical implications for enhancing user engagement and conversion strategies in e-commerce platforms, emphasizing the significance of cross-device analysis for comprehensive user behavior understanding. In Chapter 7, we examined multimodal product retrieval in e-commerce contexts, integrating textual, image, and attribute data to improve category-to-image (CTI) performance across product categories of varying granularity. We demonstrated the varying impacts of different modalities, underscoring the importance of tailoring retrieval models to leverage specific modal strengths based on product category characteristics.

In this final chapter, we revisit the main research questions raised in Chapter 1. We summarise our findings for these questions in Section 8.1. We conclude this chapter, and this thesis, with directions for future work in Section 8.2.

8.1 SUMMARY OF FINDINGS

RQ1 To what extent are the published image-text cross-modal retrieval results reproducible, replicable, and generalizable across scene-centric and object-centric datasets?

In Chapter 2, we conducted a reproducibility study involving two state-of-the-art (SOTA) CMR models, CLIP and X-VLM, evaluating their performance on both scene-centric and object-centric datasets. We focused on reproducibility, replicability, and generalizability of the results. We discovered that reproducibility of CMR results on scene-centric datasets is challenging, with partial success observed for certain tasks and metrics. We attribute the discrepancy to differences in data preprocessing and experimental setups. We found out that the replicability of relative performance from scene-centric to object-centric datasets is limited, with significant discrepancies in model performance, highlighting the importance of dataset characteristics. The generalizability of the CMR methods across different dataset types is constrained. Performance on object-centric datasets tends to be lower, suggesting that current models are not adequately robust across diverse types of data.

Thus, our response to RQ1 is that while relative performance results in image-text CMR are partially reproducible and replicable across certain datasets, particularly scene-centric ones, they face challenges on object-centric datasets. The absolute performance scores on object-centric datasets are lower compared to scene-centric datasets, emphasising the need for further exploration and evaluation of CMR methods on di-

verse benchmark datasets. Reproducibility on scene-centric datasets is a challenge, with partial success attributed to differences in data preprocessing and experimental setups. Replicability from scene-centric to object-centric datasets is limited, indicating discrepancies in model performance due to dataset characteristics. Generalizability across different dataset types is constrained, with lower performance on object-centric datasets suggesting that current models lack robustness across diverse data types.

RQ2 How can learned sparse retrieval techniques be applied in the vision-language domain?

In Chapter 3, we propose a method for multimodal learned sparse retrieval, focusing on converting dense representations into sparse ones within the VL domain. The method shows promising results in terms of both effectiveness and efficiency. We discovered that in terms of efficiency, our models achieve competitive performance compared to the original dense models. Relaxing the sparsity regularization allows the model to capture more information, leading to effectiveness closer to the dense baseline. When it comes to efficiency, our models are more efficient than dense models, requiring fewer FLOPs for retrieval. This efficiency increases with stricter sparsity regularization. Overall, we point out a trade-off between effectiveness and efficiency. As we prioritize efficiency, effectiveness slightly drops.

Consequently, we conclude for RQ2 that in the vision-language domain, learned sparse retrieval techniques can be applied by converting dense representations into sparse ones, showing promising results in both effectiveness and efficiency.

RQ3 In the context of vision-language representation learning with multiple captions per image, to what extent does the presence of a shortcut hinder learning task-optimal representations?

In Chapter 4, we explored the behaviour of contrastive learning approaches in VL in the context of shortcut learning, especially when dealing with datasets where each image has multiple captions. We proposed a novel framework, synthetic shortcuts for vision-language (SVL), to analyze this problem in a controlled setting. We discovered that contrastive losses tend to prioritize learning easily detectable features shared between the image and all captions, neglecting other relevant information that might be unique to specific captions. This dependence on shortcuts hinders the model from achieving task-optimal representations that capture the full spectrum of information within the image and its captions.

Hence, our findings for RQ3 indicate that in vision-language representation learning with multiple captions per image, the presence of shortcuts hinders the learning of task-optimal representations. We assume that this happens because contrastive learning approaches prioritize easily detectable features shared between the image and

all captions, neglecting unique information specific to individual captions. This dependence on shortcuts prevents models from capturing the full spectrum of relevant information within the image and its captions, resulting in suboptimal representations.

RQ4 How can we improve the evaluation and benchmarking of vision-language models on the image-text retrieval task?

In Chapter 5, we explored the topic of the brittleness of existing evaluation benchmarks for the ITR task. We highlight two main concerns: the granularity of existing benchmarks and the limitations of current evaluation metrics. We analyzed two ITR benchmarks and compared them with their augmented counterparts. We evaluated four SOTA VL models of the datasets. We introduced an evaluation framework that incorporates a taxonomy of perturbations designed to test the model’s robustness to changes in input data. We discovered that finer-grained datasets improve vision-language models (VLMs) performance on the task. Besides, we discovered that VLMs are sensitive to input changes: Introducing variations to the input texts generally decreased VLM performance. However, models performed better on the finer-grained datasets even with these variations. These findings highlight the importance of using more detailed datasets and robust evaluation methods to accurately assess VLMs capabilities. This will help develop VLMs that are more robust to variability in model input.

Therefore, our answer to RQ4 is that improving the evaluation and benchmarking process of vision-language models on the image-text retrieval task involves addressing the granularity of benchmarks and limitations of current evaluation metrics. Finer-grained datasets enhance model performance even when input variations are introduced, highlighting the sensitivity of models to changes in input data. An evaluation framework incorporating a taxonomy of perturbations can test model robustness, emphasizing the need for detailed datasets and robust evaluation methods to accurately assess model capabilities and develop models resilient to input variability.

RQ5 How can we facilitate product retrieval by predicting purchase intent in cross-device setting?

In Chapter 6, we analyzed user purchase intent within the realm of e-commerce. We examined user session logs spanning four weeks from a European e-commerce platform. Our analysis aimed to identify behavioral signals that would facilitate product retrieval by indicating purchase intent. We considered such factors as session duration, day of the week, session start time, device type, channel of access, and user queries. Subsequently, we explored the relevance of these identified signals through a series of experiments tailored to predict purchase intent in both anonymous and identified user settings. These experiments employed various models, including the random forest

model suited for production environments, along with five other models to ascertain the robustness and efficacy of our engineered features. Our investigation provided insights into purchasing behavior patterns within e-commerce contexts. We demonstrated that certain behavioral cues—such as session duration and timing—strongly correlate with user purchase intent. Notably, we found that these predictive insights can be leveraged early in user sessions, aiding in the anticipation of purchasing decisions. Additionally, our study underscored the challenges in discerning purchase intent among anonymous users, given the absence of prior behavioral data. Despite these challenges, our findings emphasize the significant role of device transitions in predicting purchase behavior, highlighting the necessity of cross-device analysis in e-commerce platforms.

Overall, our conclusion for RQ5 is that facilitating product retrieval by predicting purchase intent in a cross-device setting involves analyzing behavioral signals from user session logs. Factors such as session duration, timing, device type, and user queries significantly correlate with purchase intent and can be leveraged early in user sessions to anticipate purchasing decisions. Predictive insights are particularly challenging among anonymous users due to the lack of prior behavioral data. Device transitions play a crucial role in predicting purchase behavior, underscoring the necessity of cross-device analysis in e-commerce platforms.

RQ6 How do multimodal document representation, encompassing text, image, and attribute data, impact the performance on the category-to-image retrieval in the context of categories of varying granularity?

In Chapter 7, we conducted a series of experiments on an adapted e-commerce dataset containing textual descriptions, images, and attribute information of products across a diverse set of categories. We established several unimodal and bimodal baselines and implemented a multimodal retrieval model CLIP-ITA. We discovered that the multimodal document representation, which integrates text, image, and attribute data, generally improves the performance of the CTI retrieval task across categories of varying granularity. We attribute this improvement to the richer context provided by combining different modalities, which helps in accurately identifying and retrieving relevant images. In addition, the impact of multimodal representations varies with category granularity. For the most general categories, models that rely solely on image information tend to perform slightly better. This highlights the significance of visual cues in recognizing broader product categories, where detailed textual or attribute data might not be as important. However, the inclusion of textual and attribute data becomes more critical for specific categories. This suggests that the level of detail required for accurate retrieval increases with the specificity of the category.

To summarize, we answer RQ6 by saying that multimodal document representation,

integrating text, image, and attribute data, improves CTI retrieval performance across categories of varying granularity. The CLIP-ITA model, designed for this task, demonstrates superior performance by combining multiple modalities. For the most general categories, models using image information alone perform slightly better, highlighting the importance of visual cues. However, for more specific categories, the inclusion of textual and attribute data becomes important, indicating that detailed information is necessary for accurate retrieval as category specificity increases.

8.2 FUTURE WORK

The topic of multimodal machine learning in the context of information retrieval opens numerous research avenues, some of which we have touched upon throughout this thesis. We believe that these areas present interesting research opportunities. In this section, we highlight potential directions for future work that have not been explicitly addressed yet.

Dense and Sparse Retrieval We have investigated the reproducibility of CMR results on scene-centric and object-centric datasets in Chapter 2. To gain a deeper understanding of models on the CMR task, an important next step is to expand the list of scene-centric and object-centric datasets used for investigation and increase the number of models used. Investigating CMR performance on a wider range of datasets will contribute to assessing the model’s generalizability and robustness. Furthermore, exploring beyond the zero-shot scenario into few-shot and multi-shot settings will provide insights into the model’s adaptability to limited data conditions, and would allow us to gain a more comprehensive picture of the model’s capabilities on the CMR task.

Similarly, when it comes to multimodal learned sparse retrieval (MLSR) (Chapter 3), increasing the number of datasets and dense backbones as well as extending the set of modalities towards video and audio would provide a more comprehensive evaluation of the proposed model.

Representation Learning and Evaluation Building upon our work on shortcuts for contrastive VL representation learning with multiple captions per image (Chapter 4), we propose several directions for future research. Developing optimization objectives specifically tailored to address shortcut learning in this context is important for training models that rely on shortcuts less. Investigating alternative training strategies and loss functions can further improve model robustness and generalization. Additionally, combining existing shortcut reduction methods or exploring novel techniques has the potential to achieve significant performance gains. Finally, extending the SVL framework to better capture nuances and complexities of natural data is another important and promising direction. This would allow a more comprehensive exploration

of shortcut learning and the understanding of the implications in real-world scenarios and datasets.

The first next step towards improving the reliability of the ITR evaluation benchmarks (Chapter 5) would imply expanding upon the proposed framework. The expansion can go in several directions. First, developing additional methods to create more realistic text alterations. Second, evaluating a broader range of datasets and VLM will strengthen the generalizability of the findings and provide insights for the development of more robust models. Third, expanding the number of datasets and going beyond image and text data would make the framework more comprehensive. Finally, evaluating a wider range of VLM with various architectures and training methodologies would provide more generalizable insights into how these models perform under different conditions.

Product Retrieval Regarding the topic of understanding and modelling user intents across multiple devices (Chapter 6), expanding the analysis of general cross-device user search behavior within e-commerce is a promising direction. This implies investigating additional facets of user behavior and experimenting with diverse user intent modeling approaches beyond those explored in the study. Besides, further exploration of device-specific user behavior—across both commonly used devices like PCs, smartphones, and tablets, as well as less studied devices such as TVs and game consoles presents an interesting direction for further research. Such investigation can lead to more comprehensive and accurate intent prediction.

Given the investigation on CTI retrieval carried out in Chapter 7, one promising direction for future work relates to validating the robustness and generalizability of the proposed model by evaluating its performance on diverse datasets outside the e-commerce domain, such as those from digital humanities, medical image retrieval, or multimedia content analysis. Improving the model’s generalization to unseen categories is another important challenge. Besides, leveraging additional training data, advanced transfer learning techniques, or external knowledge sources can enhance the model’s ability to handle unseen concepts.

8.3 FINAL REMARKS

In this thesis we have explored multimodal machine learning for information retrieval from a vision and language perspective. We have addressed a range of challenges within the domains of dense and sparse retrieval, representation learning and evaluation, and product retrieval. By developing novel methods and evaluating existing frameworks, we have contributed to a more nuanced understanding of VLMs robustness, efficiency, and generalizability in the context of information retrieval. This journey is far from

complete, and the directions outlined in the future work open up exciting possibilities for continued research that can push the boundaries of what these models can achieve in complex and dynamic environments.

BIBLIOGRAPHY

- Aafaq, Nayyer, Naveed Akhtar, Wei Liu, Mubarak Shah, and Ajmal Mian (2021). "Controlled Caption Generation for Images Through Adversarial Attacks". In: *CoRR* abs/2107.03050.
- ACM (2020). *Artifact Review and Badging - Current*. <https://www.acm.org/publications/policies/artifact-review-and-badging-current>. Accessed August 7, 2024.
- Adnan, Mohammed, Yani Ioannou, Chuan-Yung Tsai, Angus Galloway, H.R. Tizhoosh, and Graham W. Taylor (2022). "Monitoring Shortcut Learning Using Mutual Information". In: *arXiv preprint arXiv:2206.13034*.
- Agichtein, Eugene, Eric Brill, Susan Dumais, and Robert Ragno (2006). "Learning User Interaction Models for Predicting Web Search Result Preferences". In: *Proc. SIGIR*. ACM, pp. 3–10.
- Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. (2022). "Flamingo: A Visual Language Model for Few-Shot Learning". In: *arXiv preprint arXiv:2204.14198*.
- Alberti, Chris, Jeffrey Ling, Michael Collins, and David Reitter (2019). "Fusion of Detected Objects in Text for Visual Question Answering". In: *EMNLP*, pp. 2131–2140.
- Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer Normalization". In: *arXiv preprint arXiv:1607.06450*.
- Bachman, Philip, R. Devon Hjelm, and William Buchwalter (2019). "Learning Representations by Maximizing Mutual Information Across Views". In: *NeurIPS*, pp. 15509–15519.
- Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2019). "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41, pp. 423–443.
- Bao, Hangbo, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhajit Som, Songhao Piao, and Furu Wei (2022). "VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts". In: *NeurIPS*, pp. 32897–32912.
- Bardes, Adrien, Jean Ponce, and Yann LeCun (2022). "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning". In: *ICLR*.
- Bartolo, Max, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela (2021). "Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, pp. 8830–8848.
- Bellman, Steven, Gerald Lohse, and Eric J. Johnson (1999). "Predictors of Online Buying Behavior". In: *Communications of the ACM* 42, pp. 32–48.
- Ben-Shimon, David, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle (2015). "RecSys Challenge 2015 and the YOOCHOOSE Dataset". In: *9th ACM RecSys*. ACM, pp. 357–358.
- Binder, Jeffrey R (2015). "The Wernicke Area: Modern Evidence and a Reinterpretation". In: *Neurology* 85.24, pp. 2170–2175.

- Binder, Marc D., Nobutaka Hirokawa, Uwe Windhorst, et al. (2009). *Encyclopedia of Neuroscience*. Vol. 3166. Springer Berlin, Germany.
- Biten, Ali Furkan, Andrés Mafla, Lluís Gómez, and Dimosthenis Karatzas (2022). "Is An Image Worth Five Sentences? A New Look into Semantics for Image-Text Matching". In: WACV. IEEE, pp. 2483–2492.
- Bleeker, Maurits, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke (2024). "Demonstrating and Reducing Shortcuts in Vision-Language Representation Learning". In: *Transactions on Machine Learning Research*. url: <https://openreview.net/forum?id=gfANevPraH>.
- Bleeker, Maurits, Andrew Yates, and Maarten de Rijke (2022). "Reducing Predictive Feature Suppression in Resource-Constrained Contrastive Image-Caption Retrieval". In: *arXiv preprint arXiv:2204.13382*.
- Bomhardt, Christian, Wolfgang Gaul, and Lars Schmidt-Thieme (2005). "Web Robot Detection-preprocessing Web Logfiles for Robot Detection". In: *New Developments in Classification and Data Analysis*, pp. 113–124.
- Bonab, Hamed, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan (2021). "Cross-Market Product Recommendation". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM.
- Broder, Andrei (2002). "A Taxonomy of Web Search". In: *ACM SIGIR Forum* 36.2, pp. 3–10.
- Brown, Andrew, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman (2020). "Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval". In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. Vol. 12354. Springer, pp. 677–694.
- Brown, Mark, Nigel Pope, and Kevin Voges (2003). "Buying or Browsing? An Exploration of Shopping Orientations and Online Purchase Intention". In: *European Journal of Marketing* 37.11/12, pp. 1666–1684.
- Cao, Min, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang (2022a). "Image-Text Retrieval: A Survey on Recent Research and Development". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022*. Ed. by Luc De Raedt. ijcai.org, pp. 5410–5417.
- Cao, Yu, Dianqi Li, Meng Fang, Tianyi Zhou, Jun Gao, Yibing Zhan, and Dacheng Tao (2022b). "TASA: Deceiving Question Answering Models by Twin Answer Sentences Attack". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, pp. 11975–11992.
- Carrara, Fabio, Andrea Esuli, Tiziano Fagni, Fabrizio Falchi, and Alejandro Moreo Fernández (2018). "Picture It in Your Mind: Generating High-Level Visual Representations from Textual Descriptions". In: *Information Retrieval Journal* 21, pp. 208–229.
- Carvalho, Micael, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord (2018). "Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 35–44.
- Chen, Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Madappally Jose, Alexander Toshev, Jonathon Shlens, Ruoming Pang, and Yinfei Yang (2023a). "STAIR: Learning Sparse Text and Image Representation in Grounded Tokens". In: *arXiv preprint arXiv:2301.13081*.
- Chen, Hongge, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh (2018). "Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, pp. 2587–2597.
- Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton (2020a). "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*, pp. 1597–1607.

- Chen, Ting, Calvin Luo, and Lala Li (2021). "Intriguing Properties of Contrastive Losses". In: *NeurIPS*, pp. 11834–11845.
- Chen, Weijing, Linli Yao, and Qin Jin (2023b). "Rethinking Benchmarks for Cross-Modal Image-Text Retrieval". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. Ed. by Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete. ACM, pp. 1241–1251.
- Chen, Xinlei, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick (2015). "Microsoft COCO Captions: Data Collection and Evaluation Server". In: *arXiv preprint arXiv:1504.00325*.
- Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu (2020b). "UNITER: Universal Image-Text Representation Learning". In: *ECCV*, pp. 104–120.
- Chen, Zhongzhi, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu (2023c). "AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities". In: *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics, pp. 8666–8682.
- Cheng, Justin, Caroline Lo, and Jure Leskovec (2017). "Predicting Intent Using Activity Logs: How Goal Specificity and Temporal Range Affect User Behavior". In: *Proc. WWW. International World Wide Web Conferences Steering Committee*, pp. 593–601.
- Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *ACL*, pp. 1724–1734.
- Collins, Jasmine, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik (2022). "ABO: Dataset and Benchmarks for Real-World 3D Object Understanding". In: *CVPR*.
- Dai, Honghua Kathy, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, and Ying Li (2006). "Detecting Online Commercial Intention (OCI)". In: *Proc. WWW. ACM*, pp. 829–837.
- Dai, Zhuyun and Jamie Callan (2019). "Context-Aware Sentence/Passage Term Importance Estimation for First Stage Retrieval". In: *arXiv preprint arXiv:1910.10687*.
- Dai, Zihang, Guokun Lai, Yiming Yang, and Quoc V. Le (2020). "Funnel-Transformer: Filtering Out Sequential Redundancy for Efficient Language Processing". In: *arXiv preprint arXiv:2006.03236*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT*.
- Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua 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: *ICLR*.
- Dou, Zi-Yi, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. (2022). "An Empirical Study of Training End-to-End Vision-and-Language Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18166–18176.
- Faghri, Fartash, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler (2018). "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives". In: *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, p. 12.
- Fan, Yixing, Jiafeng Guo, Xinyu Ma, Ruqing Zhang, Yanyan Lan, and Xueqi Cheng (2021). "A Linguistic Study on Relevance Modeling in Information Retrieval". In: *Proceedings of the Web Conference 2021*, pp. 1053–1064.

- Federici, Marco, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata (2020). "Learning Robust Representations via Multi-View Information Bottleneck". In: *ICLR*.
- Formal, Thibault, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant (2022). "From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery, pp. 2353–2359.
- Formal, Thibault, Benjamin Piwowarski, and Stéphane Clinchant (2021). "SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2288–2292.
- Frome, Andrea, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov (2013). "Devise: A Deep Visual-Semantic Embedding Model". In: *Advances in Neural Information Processing Systems* 26.
- Gao, Dehong, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang (2020). "Fashion-BERT: Text and Image Matching with Adaptive Loss for Cross-Modal Retrieval". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2251–2260.
- Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann (2020). "Shortcut Learning in Deep Neural Networks". In: *Nature Machine Intelligence*, pp. 665–673.
- George, Joey F. (2002). "Influences on the Intent to Make Internet Purchases". In: *Internet Research* 12.2, pp. 165–180.
- Goei, Kenneth, Mariya Hendriksen, and Maarten de Rijke (2021). "Tackling Attribute Fine-grainedness in Cross-modal Fashion Search with Multi-level Features". In: *SIGIR 2021 Workshop on eCommerce*. ACM.
- Gong, Yunchao, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik (2014). "Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections". In: *European Conference on Computer Vision*. Springer, pp. 529–545.
- Grill-Spector, Kalanit, Zoe Kourtzi, and Nancy Kanwisher (2001). "The Lateral Occipital Complex and Its Role in Object Recognition". In: *Vision Research* 41.10-11, pp. 1409–1422.
- Gu, Jiuxiang, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang (2018). "Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189.
- Guo, Long, Lifeng Hua, Rongfei Jia, Binqiang Zhao, Xiaobo Wang, and Bin Cui (2019a). "Buying or Browsing?: Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior". In: *Proc. SIGKDD*. ACM, pp. 1984–1992.
- Guo, Qi and Eugene Agichtein (2010). "Ready to Buy or Just Browsing?: Detecting Web Searcher Goals from Interaction Data". In: *Proc. SIGIR*. ACM, pp. 130–137.
- Guo, Wenzhong, Jianwen Wang, and Shiping Wang (2019b). "Deep Multimodal Representation Learning: A Survey". In: *IEEE Access* 7, pp. 63373–63394.
- Gupta, Tanmay, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem (2020). "Contrastive Learning for Weakly Supervised Phrase Grounding". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16. Springer, pp. 752–768.
- Han, Xintong, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis (2017). "Automatic Spatially-Aware Fashion Concept Discovery". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1463–1471.
- He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick (2020). "Momentum Contrast for Unsupervised Visual Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738.

- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *CVPR*, pp. 770–778.
- Hendriksen, Mariya (2022). "Multimodal Retrieval in E-Commerce: From Categories to Images, Text, and Back". In: *European Conference on Information Retrieval*. Springer, pp. 505–512.
- Hendriksen, Mariya, Maurits Bleeker, Svitlana Vakulenko, Nanne van Noord, Ernst Kuiper, and Maarten de Rijke (2022). "Extending CLIP for Category-to-image Retrieval in E-commerce". In: *ECIR 2022: 44th European Conference on Information Retrieval*. Springer, pp. 289–303.
- Hendriksen, Mariya, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke (2020). "Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers". In: *eCOM 2020: The 2020 SIGIR Workshop on eCommerce*. ACM.
- Hendriksen, Mariya, Artuur Leeuwenberg, and Marie-Francine Moens (2021). "LSTM for Dialogue Breakdown Detection: Exploration of Different Model Types and Word Embeddings". In: *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Springer, pp. 443–453.
- Hendriksen, Mariya and Viggo Overes (2022). "Unimodal vs. Multimodal Siamese Networks for Outfit Completion". In: *arXiv preprint arXiv:2207.10355*.
- Hendriksen, Mariya, Svitlana Vakulenko, Ernst Kuiper, and Maarten de Rijke (2023). "Scene-centric vs. Object-centric Image-Text Cross-modal Retrieval: A Reproducibility Study". In: *ECIR 2023: 45th European Conference on Information Retrieval*. Springer, pp. 68–85.
- Hendriksen, Mariya, Shuo Zhang, Ridho Reinanda, Mohamed Yahya, Edgar Meij, and Maarten de Rijke (2024). "Assessing Brittleness of Image-Text Retrieval Benchmarks from Vision-Language Models Perspective". In: *arXiv preprint arXiv:2407.15239*.
- Hendrycks, Dan and Kevin Gimpel (2016). "Gaussian Error Linear Units (GELUs)". In: *arXiv preprint arXiv:1606.08415*.
- Hermann, Katherine L. and Andrew K. Lampinen (2020). "What Shapes Feature Representations? Exploring Datasets, Architectures, and Training". In: *NeurIPS*, pp. 9995–10006.
- Hernández, Blanca, Julio Jiménez, and M. José Martín (2011). "Age, Gender and Income: Do They Really Moderate Online Shopping Behaviour?" In: *Online information review* 35.1, pp. 113–133.
- Herranz, Luis, Shuqiang Jiang, and Xiangyang Li (2016). "Scene Recognition with CNNs: Objects, Scales and Dataset Bias". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 571–579.
- Hessel, Jack and Alexandra Schofield (2021). "How Effective is BERT without Word Ordering? Implications for Language Understanding and Data Privacy". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 204–211.
- Hewawalpita, Supun and Indika Perera (2019). "Multimodal User Interaction Framework for E-Commerce". In: *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*. IEEE, pp. 9–16.
- Hjelm, R. Devon, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio (2019). "Learning Deep Representations by Mutual Information Estimation and Maximization". In: *ICLR*.
- Hop, Walter (2013). "Web-shop Order Prediction Using Machine Learning". MA thesis. Erasmus University Rotterdam.
- Hsu, Meng-Hsiang, Chia-Hui Yen, Chao-Min Chiu, and Chun-Ming Chang (2006). "A Longitudinal Investigation of Continued Online Shopping Behavior: An Extension of the Theory of Planned Behavior". In: *International Journal of Human-Computer Studies* 64.9, pp. 889–904.

- Hu, Peng, Liangli Zhen, Dezhong Peng, and Pei Liu (2019). "Scalable Deep Multimodal Learning for Cross-Modal Retrieval". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 635–644.
- Hu, Ronghang, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell (2016). "Natural Language Object Retrieval". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555–4564.
- Jia, Chao, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig (2021). "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision". In: *International Conference on Machine Learning*. PMLR, pp. 4904–4916.
- Jiang, Bin, Jiachen Yang, Zhihan Lv, Kun Tian, Qinggang Meng, and Yan Yan (2017). "Internet Cross-Media Retrieval Based on Deep Learning". In: *Journal of Visual Communication and Image Representation* 48, pp. 356–366.
- Jin, Di, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits (2020). "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 8018–8025.
- Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2019). "Billion-Scale Similarity Search with GPUs". In: *IEEE Transactions on Big Data* 7.3, pp. 535–547.
- Jones, K. Sparck, Steve Walker, and Stephen E. Robertson (2000). "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2". In: *Information Processing and Management* 36.6, pp. 809–840.
- Jurafsky, Dan and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Edition*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Pearson Education International.
- Kamalloo, Ehsan, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin (2023). "Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard". In: *arXiv preprint arXiv:2306.07471*.
- Karpathy, Andrej and Fei-Fei Li (2015). "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *CVPR*, pp. 3128–3137.
- Kaur, Parminder, Husanbir Singh Pannu, and Avleen Kaur Malhi (2021). "Comparative Analysis on Cross-Modal Information Retrieval: A Review". In: *Computer Science Review* 39, p. 100336.
- Kaushik, Divyansh, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih (2021). "On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics, pp. 6618–6633.
- Kim, Eunju, Wooju Kim, and Yillbyung Lee (2003). "Combination of Multiple Classifiers for the Customer's Purchase Behavior Prediction". In: *Decision Support Systems* 34.2, pp. 167–175.
- Kim, Wonjae, Bokyung Son, and Ildoo Kim (2021). "Vilt: Vision-and-Language Transformer without Convolution or Region Supervision". In: *International Conference on Machine Learning*. PMLR, pp. 5583–5594.
- Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *ICLR*.
- Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *ICLR*.
- Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel (2014). "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models". In: *arXiv preprint arXiv:1411.2539*.
- Klein, Benjamin, Guy Lev, Gil Sadeh, and Lior Wolf (2014). "Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation". In: *arXiv preprint arXiv:1411.7399*.

- Kondylidis, Nikolaos, Jie Zou, and Evangelos Kanoulas (2021). "Category Aware Explainable Conversational Recommendation". In: *arXiv preprint arXiv:2103.08733*.
- Kovatchev, Venelin, Trina Chatterjee, Venkata Subrahmanyam Govindarajan, Jifan Chen, Eunsol Choi, Gabriella Chronis, Anubrata Das, Katrin Erk, Matthew Lease, Junyi Jessy Li, Yating Wu, and Kyle Mahowald (2022). "Longhorns at DADC 2022: How Many Linguists Does It Take to Fool a Question Answering Model? A Systematic Approach to Adversarial Attacks". In: *CoRR abs/2206.14729*.
- Krause, Jonathan, Michael Stark, Jia Deng, and Li Fei-Fei (2013). "3D Object Representations for Fine-Grained Categorization". In: *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia.
- Laenen, Katrien (2022). "Cross-Modal Representation Learning for Fashion Search and Recommendation". PhD thesis. KU Leuven.
- Laenen, Katrien and Marie-Francine Moens (2019). "Multimodal Neural Machine Translation of Fashion E-Commerce Descriptions". In: *International Conference on Fashion communication: between tradition and future digital developments*. Springer, pp. 46–57.
- Laenen, Katrien and Marie-Francine Moens (2020). "A Comparative Study of Outfit Recommendation Methods with a Focus on Attention-based Fusion". In: *Information Processing & Management* 57.6, p. 102316.
- Laenen, Katrien, Susana Zoghbi, and Marie-Francine Moens (2017). "Cross-Modal Search for Fashion Attributes". In: *Proceedings of the KDD 2017 Workshop on Machine Learning Meets Fashion*. Vol. 2017. ACM, pp. 1–10.
- Laenen, Katrien, Susana Zoghbi, and Marie-Francine Moens (2018). "Web Search of Fashion Items with Multimodal Querying". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 342–350.
- Lang, Tobias and Matthias Rettenmeier (2017). "Understanding Consumer Behavior with Recurrent Neural Networks". In: *Workshop on Machine Learning Methods for Recommender Systems*.
- Lee, Kuang-Huei, Anurag Arnab, Sergio Guadarrama, John F. Canny, and Ian Fischer (2021). "Compressive Visual Representations". In: *NeurIPS*, pp. 19538–19552.
- Lee, Kuang-Huei, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He (2018). "Stacked Cross Attention for Image-Text Matching". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216.
- Lee, Munyoung, Taehoon Ha, Jinyoung Han, Jong-Youn Rha, and Ted Taekyoung Kwon (2015). "Online Footsteps to Purchase: Exploring Consumer Behaviors on Online Shopping Sites". In: *Proc. WebSci*, pp. 1–10.
- Lewis, Molly L. and Michael C. Frank (2016). "The Length of Words Reflects Their Conceptual Complexity". In: *Cognition* 153, pp. 182–195.
- Li, Ang, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten (2017). "Learning Visual N-Grams from Web Data". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4183–4192.
- Li, Dongxu, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi (2022a). "LAVIS: A Library for Language-Vision Intelligence". In: *arXiv preprint arXiv:2209.09019*.
- Li, Gen, Nan Duan, Yuejian Fang, Dixin Jiang, and Ming Zhou (2020a). "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training". In: *AAAI*.
- Li, Haoran, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou (2020b). "Aspect-Aware Multimodal Summarization for Chinese E-commerce Products". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8188–8195.
- Li, Jinfeng, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang (2018). "TextBugger: Generating Adversarial Text Against Real-World Applications". In: *arXiv preprint arXiv:1812.05271*.

- Li, Junnan, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi (2023a). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". In: *ICML*, pp. 19730–19742.
- Li, Junnan, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi (2022b). "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *ICML*, pp. 12888–12900.
- Li, Junnan, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi (2021a). "Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation". In: *NeurIPS*, pp. 9694–9705.
- Li, Kunpeng, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu (2019a). "Visual Semantic Reasoning for Image-Text Matching". In: *ICCV*, pp. 4654–4662.
- Li, Linjie, Jie Lei, Zhe Gan, and Jingjing Liu (2021b). "Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models". In: *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, pp. 2022–2031.
- Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang (2019b). "VisualBERT: A Simple and Performant Baseline for Vision and Language". In: *arXiv preprint arXiv:1908.03557*.
- Li, Tianhong, Lijie Fan, Yuan Yuan, Hao He, Yonglong Tian, Rogério Feris, Piotr Indyk, and Dina Katabi (2023b). "Addressing Feature Suppression in Unsupervised Visual Representations". In: *WACV*, pp. 1411–1420.
- Li, Xingchen, Xiang Wang, Xiangnan He, Long Chen, Jun Xiao, and Tat-Seng Chua (2020c). "Hierarchical Fashion Graph Network for Personalized Outfit Recommendation". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 159–168.
- Li, Xiujun, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. (2020d). "OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks". In: *ECCV*, pp. 121–137.
- Li, Yangguang, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan (2022c). "Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm". In: *ICLR*.
- Li, Yehao, Jiahao Fan, Yingwei Pan, Ting Yao, Weiyao Lin, and Tao Mei (2022d). "Uni-EDEN: Universal Encoder-Decoder Network by Multi-Granular Vision-Language Pre-Training". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18.2, pp. 1–16.
- Liang, Paul Pu, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Russ Salakhutdinov (2023). "Factorized Contrastive Learning: Going Beyond Multi-view Redundancy". In: *NeurIPS*.
- Liao, Lizi, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua (2018). "Interpretable Multimodal Retrieval for Fashion Products". In: *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1571–1579.
- Lin, Jimmy and Xueguang Ma (2021). "A Few Brief Notes on DeepImpact, Coil, and a Conceptual Framework for Information Retrieval Techniques". In: *arXiv preprint arXiv:2106.14807*.
- Lin, Sheng-Chieh and Jimmy Lin (2021). "Densifying Sparse Representations for Passage Retrieval by Representational Slicing". In: *arXiv preprint arXiv:2112.04666*.
- Lin, Sheng-Chieh and Jimmy Lin (2023). "A Dense Representation Framework for Lexical and Semantic Matching". In: *ACM Transactions on Information Systems* 41.4, pp. 1–29.
- Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick (2014). "Microsoft COCO: Common Objects in Context". In: *ECCV*, pp. 740–755.
- Lin, Yujie, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke (2019). "Improving Outfit Recommendation with Co-Supervision of Fashion Generation". In: *The World Wide Web Conference*, pp. 1095–1105.

- Liu, Chunxiao, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang (2019). "Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching". In: *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 3–11.
- Liu, Yu-An, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke (2024a). "Robust Information Retrieval". In: *SIGIR 2024: 47th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 3009–3012.
- Liu, Yu-An, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng (2023). "Black-Box Adversarial Attacks against Dense Retrieval Models: A Multi-View Contrastive Learning Method". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1647–1656.
- Liu, Yu-An, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng (2024b). "Multi-Granular Adversarial Attacks against Black-box Neural Ranking Models". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1391–1400.
- Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo (2021). "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.
- Lo, Caroline, Dan Frankowski, and Jure Leskovec (2016). "Understanding Behaviors that Lead to Purchasing: A Case Study of Pinterest". In: *Proc. SIGKDD*. ACM, pp. 531–540.
- Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization". In: *ICLR*.
- Lu, Haoyu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen (2022). "COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Modal Retrieval". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, pp. 15671–15680.
- Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee (2019). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *NeurIPS*, pp. 13–23.
- Luccioni, Alexandra Sasha and Alex Hernandez-Garcia (2023). "Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning". In: *arXiv preprint arXiv:2302.08476*.
- Lupart, Simon and Stéphane Clinchant (2023). "A study on FGSM adversarial training for neural retrieval". In: *European Conference on Information Retrieval*. Springer, pp. 484–492.
- MacAvaney, Sean and Craig Macdonald (2022). "A Python Interface to PISA!" In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- MacAvaney, Sean, Craig Macdonald, and Iadh Ounis (2022). "Streamlining Evaluation with ir-measures". In: *European Conference on Information Retrieval*. Springer, pp. 305–310.
- MacAvaney, Sean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellootto, Nazli Goharian, and Ophir Frieder (2020). "Expansion via Prediction of Importance with Contextualization". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1573–1576.
- Mackenzie, Joel, Andrew Trotman, and Jimmy Lin (2021). "Wacky weights in learned sparse representations and the revenge of score-at-a-time query evaluation". In: *arXiv preprint arXiv:2110.11540*.
- Mallia, Antonio, Michal Siedlaczek, Joel Mackenzie, and Torsten Suel (2019). "PISA: Performant Indexes and Search for Academia". In: *Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019*. Pp. 50–56.
- Marin, Javier, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba (2021). "Recipe1m+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1, pp. 187–203.

- McGurk, Harry and John MacDonald (1976). "Hearing Lips and Seeing Voices". In: *Nature* 264:5588, pp. 746–748.
- Messina, Nicola, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet (2021). "Fine-Grained Visual Textual Alignment for Cross-Modal Retrieval Using Transformer Encoders". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17.4, pp. 1–23.
- Moe, Wendy W. and Peter S. Fader (2004). "Dynamic Conversion Behavior at E-commerce Sites". In: *Management Science* 50.3, pp. 326–335.
- Montanez, George D., Ryen W. White, and Xiao Huang (2014). "Cross-Device Search". In: *Proc. CIKM*. ACM, pp. 1669–1678.
- Morwitz, Vicki G. and David Schmittlein (1992). "Using Segmentation to Improve Sales Forecasts Based on Purchase Intent: Which "Intenders" Actually Buy?" In: *Journal of marketing research* 29.4, pp. 391–405.
- Mu, Norman, Alexander Kirillov, David A. Wagner, and Saining Xie (2022). "SLIP: Self-Supervision Meets Language-Image Pre-training". In: *ECCV*, pp. 529–544.
- Nagarajan, Tushar and Kristen Grauman (2018). "Attributes as Operators: Factorizing Unseen Attribute-Object Compositions". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 169–185.
- Nam, Hyeonseob, Jung-Woo Ha, and Jeonghee Kim (2017). "Dual Attention Networks for Multimodal Reasoning and Matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 299–307.
- Nguyen, Thong, Mariya Hendriksen, and Andrew Yates (2023a). "Multimodal Learned Sparse Retrieval for Image Suggestion Task". In: *TREC*.
- Nguyen, Thong, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke (2024). "Multi-Modal Learned Sparse Retrieval with Probabilistic Expansion Control". In: *ECIR 2024: 46th European Conference on Information Retrieval*. Springer.
- Nguyen, Thong, Sean MacAvaney, and Andrew Yates (2023b). "A Unified Framework for Learned Sparse Retrieval". In: *European Conference on Information Retrieval*. Springer, pp. 101–116.
- Nguyen, Thong, Sean MacAvaney, and Andrew Yates (2023c). "Adapting Learned Sparse Retrieval for Long Documents". In: *arXiv preprint arXiv:2305.18494*.
- Nielsen, Jakob, Rolf Molich, Carolyn Snyder, and Susan Farrell (2000). "E-commerce User Experience". In: *Nielsen Norman Group*.
- Nilsback, Maria-Elena and Andrew Zisserman (2008). "Automated Flower Classification over a Large Number of Classes". In: *Indian Conference on Computer Vision, Graphics and Image Processing*.
- Niu, Xi, Chuqin Li, and Xing Yu (2017). "Predictive Analytics of E-Commerce Search Behavior for Conversion". In: *23rd Americas Conference on Information Systems, AMCIS 2017, Boston, MA, USA, August 10-12, 2017*. Association for Information Systems.
- O'cass, Aron and Tino Fenech (2003). "Web Retailing Adoption: Exploring the Nature of Internet Users Web Retailing Behaviour". In: *Journal of Retailing and Consumer services* 10.2, pp. 81–94.
- Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding". In: *arXiv preprint arXiv:1807.03748*.
- O'Connor, Joe and Jacob Andreas (2021). "What Context Features Can Transformer Language Models Use?" In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 851–864.
- Parry, Andrew, Maik Fröbe, Sean MacAvaney, Martin Potthast, and Matthias Hagen (2024). "Analyzing Adversarial Attacks on Sequence-to-Sequence Relevance Models". In: *European Conference on Information Retrieval*. Springer, pp. 286–302.

- Partee, Barbara (1995). "Lexical Semantics and Compositionality". In: *An Invitation to Cognitive Science: Language 1*, pp. 311–360.
- Penha, Gustavo, Arthur Câmara, and Claudia Hauff (2022). "Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators". In: *European conference on information retrieval*. Springer, pp. 397–412.
- Pesahov, Leon, Ayal Klein, and Ido Dagan (2023). "QA-Adj: Adding Adjectives to QA-based Semantics". In: *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pp. 74–88.
- Petrov, Aleksandr and Craig Macdonald (2022). "A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation". In: *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 436–447.
- Pham, Thang, Trung Bui, Long Mai, and Anh Nguyen (2021). "Out of Order: How Important is the Sequential Order of Words in a Sentence in Natural Language Understanding Tasks?" In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1145–1160.
- Piasecki, Maciej, Bernd Broda, and Stanislaw Szpakowicz (2009). *A WordNet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.
- Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik (2015). "Flickr30K Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649.
- Qi, Di, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti (2020). "ImageBERT: Cross-Modal Pre-training with Large-Scale Weak-Supervised Image-Text Data". In: *arXiv preprint arXiv:2001.07966*.
- Qin, Zengchang, Jing Yu, Yonghui Cong, and Tao Wan (2016). "Topic Correlation Model for Cross-Modal Multimedia Information Retrieval". In: *Pattern Analysis and Applications* 19, pp. 1007–1022.
- Radford, Alec, 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: *ICML*, pp. 8748–8763.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language Models are Unsupervised Multitask Learners". In: *OpenAI blog*, p. 9.
- Ram, Ori, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson (July 2023). "What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 2481–2498.
- Rao, Jun, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao (2022). "Where Does the Performance Improvement Come From? A Reproducibility Concern about Image-Text Retrieval". In: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, pp. 2727–2737.
- Reed, Scott, Zeynep Akata, Honglak Lee, and Bernt Schiele (2016). "Learning Deep Representations of Fine-Grained Visual Descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49–58.
- Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *EMNLP-IJCNLP*, pp. 3980–3990.
- Robinson, Joshua, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra (2021). "Can Contrastive Learning avoid Shortcut Solutions?" In: *NeurIPS*, pp. 4974–4986.
- Rowley, Jennifer (2000). "Product Search in E-Shopping: A Review and Research Propositions". In: *Journal of Consumer Marketing* 17.1, pp. 20–35.
- Salisbury, W. David, Rodney A. Pearson, Allison W. Pearson, and David W. Miller (2001). "Perceived Security and World Wide Web Purchase Intention". In: *Industrial Management & Data Systems* 101.4, pp. 165–177.

- Scimeca, Luca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun (2022). "Which Shortcut Cues Will DNNs Choose? A Study from the Parameter-Space Perspective". In: *ICLR*.
- Seippel, Hannah Sophia (2018). "Customer Purchase Prediction through Machine Learning". MA thesis. University of Twente.
- Senecal, Sylvain, Paweł J. Kalczynski, and Jacques Nantel (2005). "Consumers' Decision-Making Process and their Online Shopping Behavior: A Clickstream Analysis". In: *Journal of Business Research* 58.11, pp. 1599–1608.
- Sharma, Piyush, 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)*, pp. 2556–2565.
- Shen, Sheng, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer (2021). "How Much Can CLIP Benefit Vision-and-Language Tasks?" In: *arXiv preprint arXiv:2107.06383*.
- Shen, Zong-Ying, Shiang-Yu Han, Li-Chen Fu, Pei-Yung Hsiao, Yo-Chung Lau, and Sheng-Jen Chang (2019). "Deep Convolution Neural Network with Scene-Centric and Object-Centric Information for Object Detection". In: *Image and Vision Computing* 85, pp. 14–25.
- Sheng, Sasha, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela (2021a). "Human-Adversarial Visual Question Answering". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, pp. 20346–20359.
- Sheng, Shurong, Katrien Laenen, Luc Van Gool, and Marie-Francine Moens (2021b). "Fine-Grained Cross-Modal Retrieval for Cultural Items with Focal Attention and Hierarchical Encodings". In: *Computers* 10.9, p. 105.
- Shwartz-Ziv, Ravid and Yann LeCun (2023). "To Compress or Not to Compress – Self-Supervised Learning and Information Theory: A Review". In: *arXiv preprint arXiv:2304.09355*.
- Sidiropoulos, Georgios and Evangelos Kanoulas (2022). "Analyzing the Robustness of Dual Encoders for Dense Retrieval Against Misspellings". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2132–2136.
- Simonyan, Karen and Andrew Zisserman (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR*.
- Sismeiro, Catarina and Randolph E. Bucklin (2004). "Modeling Purchase Behavior at an E-commerce Web Site: A Task-completion Approach". In: *Journal of marketing research* 41.3, pp. 306–323.
- Smeulders, Arnold W.M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain (2000). "Content-Based Image Retrieval at the End of the Early Years". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.12, pp. 1349–1380.
- Song, Juyong and Sunghyun Choi (2021). "Image-Text Alignment using Adaptive Cross-attention with Transformer Encoder for Scene Graphs". In: p. 343.
- Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu (2020). "MPNet: Masked and Permuted Pre-training for Language Understanding". In: *NeurIPS*, pp. 16857–16867.
- Sridharan, Karthik and Sham M. Kakade (2008). "An Information Theoretic Framework for Multi-view Learning". In: *COLT*, pp. 403–414.
- Su, Ning, Jiayin He, Yiqun Liu, Min Zhang, and Shaoping Ma (2018). "User Intent, Behaviour, and Perceived Satisfaction in Product Search". In: *Proc. WSDM*. ACM, pp. 547–555.
- Su, Weijie, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai (2020). "VL-BERT: Pre-training of Generic Visual-Linguistic Representations". In: *ICLR*.

- Suchacka, Grażyna, Magdalena Skolimowska-Kulig, and Aneta Potempa (2015). "A k-Nearest Neighbors Method for Classifying User Sessions in E-commerce Scenario". In: *Journal of Telecommunications and Information Technology*.
- Suh, Euiho, Seungjae Lim, Hyunseok Hwang, and Suyeon Kim (2004). "A Prediction Model for the Purchase Probability of Anonymous Customers to Support Real Time Web Marketing: A Case Study". In: *Expert Systems with Applications* 27.2, pp. 245–255.
- Swinyard, William R. and Scott M. Smith (2004). "Activities, Interests, and Opinions of Online Shoppers and Non-shoppers". In: *IBER* 3.4.
- Tagliabue, Jacopo, Bingqing Yu, and Marie Beaulieu (2020). "How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead". In: *arXiv preprint arXiv:2005.12781*.
- Tan, Hao and Mohit Bansal (2019). "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *EMNLP-IJCNLP*, pp. 5099–5110.
- Tan, Mingxing and Quoc Le (2019). "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks". In: *International conference on machine learning*. PMLR, pp. 6105–6114.
- Tao, Zhiqiang, Sheng Li, Zhaowen Wang, Chen Fang, Longqi Yang, Handong Zhao, and Yun Fu (2019). "Log2Intent: Towards Interpretable User Modeling via Recurrent Semantics Memory Unit". In: *Proc. SIGKDD*. ACM, pp. 1055–1063.
- Tautkute, Ivona, Tomasz Trzciński, Aleksander P. Skorupa, Łukasz Brocki, and Krzysztof Marasek (2019). "DeepStyle: Multimodal Search Engine for Fashion and Interior Design". In: *IEEE Access* 7, pp. 84613–84628.
- Thomas, Christopher and Adriana Kovashka (2020). "Preserving Semantic Neighborhoods for Robust Cross-Modal Retrieval". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII* 16, pp. 317–335.
- Thomee, Bart, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li (2016). "YFCC100M: The New Data in Multimedia Research". In: *Communications of the ACM* 59.2, pp. 64–73.
- Tian, Yonglong, Dilip Krishnan, and Phillip Isola (2020a). "Contrastive Multiview Coding". In: *ECCV*, pp. 776–794.
- Tian, Yonglong, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola (2020b). "What Makes for Good Views for Contrastive Learning?" In: *NeurIPS*, pp. 6827–6839.
- Tomasello, Rosario, Max Garagnani, Thomas Wennekers, and Friedemann Pulvermüller (2017). "Brain Connections of Words, Perceptions and Actions: A Neurobiological Model of Spatio-Temporal Semantic Activation in the Human Cortex". In: *Neuropsychologia* 98, pp. 111–129.
- Tsagkias, Manos, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke (2020). "Challenges and Research Opportunities in eCommerce Search and Recommendations". In: *SIGIR Forum* 54.1.
- Tsai, Yao-Hung Hubert, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency (2021). "Self-Supervised Learning from a Multi-view Perspective". In: *ICLR*.
- Tschannen, Michael, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic (2020). "On Mutual Information Maximization for Representation Learning". In: *ICLR*.
- Tschannen, Michael, Manoj Kumar, Andreas Peter Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer (2023). "Image Captioners Are Scalable Vision Learners Too". In: *NeurIPS*.
- Ueki, Kazuya (2021). "Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval". In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 628–634.
- Varamesh, Ali, Ali Diba, Tinne Tuytelaars, and Luc Van Gool (2020). "Self-Supervised Ranking for Representation Learning". In: *arXiv preprint arXiv:2010.07258*.

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems 30*.
- Veldkamp, Karel, Mariya Hendriksen, Zoltán Szlávik, and Alexander Keijser (2023). "Towards Contrastive Learning in Music Video Domain". In: *arXiv preprint arXiv:2309.00347*.
- Vo, Nam, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays (2019). "Composing Text and Image for Image Retrieval: An Empirical Odyssey". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6439–6448.
- Voorhees, Ellen M. (2002). "The Philosophy of Information Retrieval Evaluation". In: *Evaluation of Cross-Language Information Retrieval Systems*. Springer Berlin Heidelberg, pp. 355–370.
- Wallace, Eric, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber (2019). "Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering". In: *Transactions of the Association for Computational Linguistics 7*. Ed. by Lillian Lee, Mark Johnson, Brian Roark, and Ani Nenkova.
- Wang, Boxin, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li (2021a). "Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*. Ed. by Joaquin Vanschoren and Sai-Kit Yeung.
- Wang, Guanhua, Hua Ji, Dexin Kong, and Na Zhang (2020). "Modality-Dependent Cross-Modal Retrieval Based on Graph Regularization". In: *Mobile Information Systems 2020*, pp. 1–17.
- Wang, Hao, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and Steven C.H. Hoi (2021b). "Cross-Modal Food Retrieval: Learning a Joint Embedding of Food Images and Recipes with Semantic Consistency and Attention Mechanism". In: *IEEE Transactions on Multimedia 24*, pp. 2515–2525.
- Wang, Haoqing, Xun Guo, Zhi-Hong Deng, and Yan Lu (2022a). "Rethinking Minimal Sufficient Representation in Contrastive Learning". In: *CVPR*, pp. 16020–16029.
- Wang, Kaiye, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang (2016a). "A Comprehensive Survey on Cross-Modal Retrieval". In: *arXiv preprint arXiv:1607.06215*.
- Wang, Liwei, Yin Li, and Svetlana Lazebnik (2016b). "Learning Deep Structure-Preserving Image-Text Embeddings". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5005–5013.
- Wang, Shuai, Shengyao Zhuang, and Guido Zuccon (2021c). "BERT-based Dense Retriever Require Interpolation with BM25 for Effective Passage Retrieval". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 317–324.
- Wang, Wenhui, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singh, Subhajit Som, et al. (2022b). "Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks". In: *arXiv preprint arXiv:2208.10442*.
- Wang, Yabing, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang (2022c). "Cross-Lingual Cross-Modal Retrieval with Noise-Robust Learning". In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 422–433.
- Wang, Yanfei, Fei Wu, Jun Song, Xi Li, and Yuetong Zhuang (2014). "Multi-Modal Mutual Topic Reinforce Modeling for Cross-Media Retrieval". In: *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 307–316.
- Welinder, P., S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona (2010). *Caltech-UCL Birds 200*. Tech. rep. CNS-TR-2010-001. California Institute of Technology.
- Wen, Keyu, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao (2021). "COOKIE: Contrastive Cross-Modal Knowledge Sharing Pre-training for Vision-Language Representation". In: *2021*

- IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021.* IEEE, pp. 2188–2197.
- Wen, Yu-Ting, Pei-Wen Yeh, Tzu-Hao Tsai, Wen-Chih Peng, and Hong-Han Shuai (2018). “Customer Purchase Behavior Prediction from Payment Datasets”. In: *Proc. WSDM*. ACM, pp. 628–636.
- Wiles, Olivia, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil (2022). “A Fine-Grained Analysis on Distribution Shift”. In: *ICLR*.
- Wirojwatanakul, Pasawee and Artit Wangperawong (2019). “Multi-Label Product Categorization Using Multi-Modal Fusion Models”. In: *arXiv preprint arXiv:1907.00420*.
- Xiao, Tete, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell (2021). “What Should Not Be Contrastive in Contrastive Learning”. In: *ICLR*.
- Xu, Gongwen, Xiaomei Li, Lin Shi, Zhijun Zhang, and Aidong Zhai (2020). “Combination Subspace Graph Learning for Cross-Modal Retrieval”. In: *Alexandria Engineering Journal* 59.3, pp. 1333–1343.
- Xu, Jiarui, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang (2022). “GroupViT: Semantic Segmentation Emerges from Text Supervision”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144.
- Xu, Xiaojun, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song (2018). “Fooling Vision and Language Models Despite Localization and Attention Mechanism”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*. Computer Vision Foundation / IEEE Computer Society, pp. 4951–4961.
- Xu, Yan, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu (2019). “Exact Adversarial Attack to Image Captioning via Structured Output Learning with Latent Variables”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4135–4144.
- Xu, Zhenlin, Yi Zhu, Tiffany Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo (2023). “Challenges of Zero-Shot Recognition with Vision-Language Models: Granularity and Correctness”. In: *arXiv preprint arXiv:2306.16048*.
- Yamaura, Yusuke, Nobuya Kanemaki, and Yukihiro Tsuboshita (2019). “The Resale Price Prediction of Secondhand Jewelry Items Using a Multi-modal Deep Model with Iterative Co-Attention”. In: *arXiv preprint arXiv:1907.00661*.
- Yang, Xun, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua (2019). “Interpretable Fashion Matching with Rich Attributes”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 775–784.
- Yao, Lewei, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu (2022). “FILIP: Fine-grained Interactive Language-Image Pre-Training”. In: *ICLR*.
- Yashima, Takuya, Naoaki Okazaki, Kentaro Inui, Kota Yamaguchi, and Takayuki Okatani (2016). “Learning to Describe E-commerce Images from Noisy Online Data”. In: *Asian Conference on Computer Vision*. Springer, pp. 85–100.
- Yim, Jonghwa, Junghun James Kim, and Daekyu Shin (2018). “One-Shot Item Search with Multimodal Data”. In: *arXiv preprint arXiv:1811.10969*.
- Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier (2014). “From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions”. In: *Transactions of the Association for Computational Linguistics* 2, pp. 67–78.
- Young Kim, Eun and Youn-Kyung Kim (2004). “Predicting Online Purchase Intentions for Clothing Products”. In: *European Journal of Marketing* 38.7, pp. 883–897.
- Yu, Jiahui, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu (2022). “Coca: Contrastive Captioners are Image-Text Foundation Models”. In: *arXiv preprint arXiv:2205.01917*.

- Yuan, Lu, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. (2021). "Florence: A New Foundation Model for Computer Vision". In: *arXiv preprint arXiv:2111.11432*.
- Yuksekgonul, Mert, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou (2023). "When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?" In: *ICLR*.
- Zamani, Hamed, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps (2018). "From Neural Re-ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 497–506.
- Zeng, Yan, Xinsong Zhang, and Hang Li (2022). "Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 25994–26009.
- Zhang, Cheng, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao (2021a). "MosaicOS: A Simple and Effective Use of Object-Centric Images for Long-Tailed Object Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 417–427.
- Zhang, Jiaming, Qi Yi, and Jitao Sang (2022a). "Towards Adversarial Attack on Vision-Language Pre-Training Models". In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013.
- Zhang, Kun, Zhendong Mao, Quan Wang, and Yongdong Zhang (2022b). "Negative-Aware Attention Framework for Image-Text Matching". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15661–15670.
- Zhang, Pengchuan, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao (2021b). "ViNVL: Revisiting Visual Representations in Vision-Language Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588.
- Zhang, Yuhao, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz (2022c). "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2022, 5-6 August 2022, Durham, NC, USA*. Vol. 182. Proceedings of Machine Learning Research. PMLR, pp. 2–25. URL: <https://proceedings.mlr.press/v182/zhang22a.html>.
- Zhao, Fei, Zhen Wu, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen (2022). "Learning from Adjective-Noun Pairs: A Knowledge-Enhanced Framework for Target-Oriented Multimodal Sentiment Classification". In: *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6784–6794.
- Zhao, Jing, Xijiong Xie, Xin Xu, and Shiliang Sun (2017). "Multi-View Learning Overview: Recent Progress and New Challenges". In: *Inf. Fusion* 38, pp. 43–54.
- Zhao, Pu, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, and Dixin Jiang (2023a). "LexLIP: Lexicon-Bottlenecked Language-Image Pre-Training for Large-Scale Image-Text Retrieval". In: *arXiv preprint arXiv:2302.02908*.
- Zhao, Yunqing, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin (2023b). "On Evaluating Adversarial Robustness of Large Vision-Language Models". In: *CoRR*.
- Zhong, Fangming, Guangze Wang, Zhikui Chen, Feng Xia, and Geyong Min (2020). "Cross-Modal Retrieval for CPSS Data". In: *IEEE Access* 8, pp. 16689–16701.
- Zhou, Bolei, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva (2014). "Learning Deep Features for Scene Recognition Using Places Database". In: *Advances in Neural Information Processing Systems* 27.

- Zhuang, Shengyao and Guido Zuccon (2022). "CharacterBERT and Self-Teaching for Improving the Robustness of Dense Retrievers on Queries with Typos". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1444–1454.
- Zhuge, Mingchen, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao (2021). "Kaleido-BERT: Vision-Language Pre-training on Fashion Domain". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12647–12657.
- Zoghbi, Susana, Geert Heyman, Juan Carlos Gomez, and Marie-Francine Moens (2016). "Cross-Modal Fashion Search". In: *International Conference on Multimedia Modeling*. Springer, pp. 367–373.
- Zong, Yongshuo, Oisin Mac Aodha, and Timothy Hospedales (2023). "Self-Supervised Multimodal Learning: A Survey". In: *arXiv preprint arXiv:2304.01008*.

SUMMARY

In this thesis, we focus on multimodal machine learning within the context of information retrieval as the main task. We focus on vision and language as core modalities. The investigation centers around three primary areas: (i) dense and sparse retrieval, (ii) representation learning and evaluation, and (iii) product retrieval. Each of the chapters in the thesis is driven by one or more of these themes.

In Chapter 2, we investigate the reproducibility of image-text cross-modal retrieval models across scene-centric and object-centric datasets. While most research focuses on scene-centric data, we examine the less explored object-centric domain. By evaluating state-of-the-art cross-modal retrieval models on both dataset types, we identify challenges in replicating results on object-centric data. This underscores the need for standardized benchmarks and methodologies to improve consistency in cross-modal retrieval research.

In Chapter 3, we focus on multimodal learned sparse retrieval and explore how sparsifying dense vectors affect model performance on the image-text retrieval task. We identify the phenomena of dimension co-activation and semantic deviation and propose metrics to quantify them. We propose a method of transforming dense vision-language representations into sparse ones and demonstrate that our approach maintains competitiveness with dense models while improving computational efficiency. We demonstrate how controlled expansion can mitigate dimension co-activation and semantic deviation, contributing to a better understanding of sparsification in multimodal retrieval.

In Chapter 4, we investigate shortcut learning in vision-language representation learning with multiple captions per image. We define shortcut learning as the use of easily detectable features that do not fully represent the task's information. We introduce the framework for synthetic shortcuts in vision-language models to analyze the reliance on synthetic shortcuts during training and evaluation, showing that contrastive vision-language methods often depend on shortcuts, neglecting all task-relevant information. The study underscores the need to address shortcut learning to improve the robustness of vision-language representation learning.

In Chapter 5, we focus on improving the evaluation of vision-language models on the image-text retrieval task in the context of concept granularity. We analyze the concept granularity of existing benchmarks and propose a novel evaluation framework

that comprises a taxonomy of perturbations and a cross-modal evaluation metric. We evaluate four state-of-the-art vision-language models on the benchmark, investigate the impact of concept granularity on performance, and advocate for a refined evaluation pipeline that better reflects real-world complexities.

In Chapter 6, we tackle the problem of improving product retrieval by predicting purchase intent in cross-device scenarios, distinguishing between anonymous and identified sessions. We analyze session logs from a European e-commerce platform to identify purchase intent signals and develop predictive models that consider session-based and historical features. We demonstrate that purchase intent can be predicted early in the session, with users often switching to devices with screens for final purchases. This work contributes to our understanding of user behaviour across devices for both anonymous and identified sessions.

Finally, in Chapter 7, we propose and motivate category-to-image retrieval task and explore the impact of multimodal product representations on this task. We combine textual, visual, and attribute information, and investigate their impact on performance in the context of categories of varying granularity. Experiments demonstrate that multimodal representations generally improve performance, although image-only models can outperform multimodal ones for general categories. The findings highlight the importance of considering the interplay between different modalities when building retrieval models.

SAMENVATTING

In dit proefschrift richten we ons op multimodale machine learning binnen de context van informatieophaling als hoofdtaak. We concentreren ons op visie en taal als kern-modaliteiten. Het onderzoek draait om drie primaire gebieden: (i) *dense and sparse retrieval*, (ii) *representation learning and evaluation*, en (iii) *product retrieval*. In elk van de hoofdstukken van dit proefschrift wordt één van deze thema's behandeld.

In Hoofdstuk 2 onderzoeken we de reproduceerbaarheid van *image-text cross-modal retrieval* modellen over scène-centrische en object-centrische datasets. Terwijl het meeste onderzoek zich richt op scène-centrische data, bekijken wij het minder verkende object-centrische domein. Door *state-of-the-art cross-modal retrieval* modellen op beide datasettypes te evalueren, identificeren we problemen bij het repliceren van resultaten op object-centrische data. Dit benadrukt de noodzaak van gestandaardiseerde *benchmarks* en methodologieën om de consistentie in *cross-modal retrieval* onderzoek te verbeteren.

In Hoofdstuk 3 richten we ons op *multimodal learned sparse retrieval* en onderzoeken we hoe het uittunnen van dichte vectoren de modelprestaties op de *image-text retrieval* taak beïnvloedt. We identificeren het fenomeen van dimensie-coactivatie en semantische afwijking, en stellen metrieken voor om deze te kwantificeren. We stellen een methode voor om dichte *vision-language* representaties om te zetten in dunne representaties en laten zien dat onze benadering de goede concurrentiepositie met dichte modellen behoudt terwijl de rekenefficiëntie verbetert. We tonen aan hoe gecontroleerde expansie dimensie-coactivatie en semantische afwijking kan verminderen, wat bijdraagt aan een beter begrip van uitdunning in multimodale ophaling.

In Hoofdstuk 4 onderzoeken we *shortcut learning* in *vision-language representation learning* met meerdere labels per afbeelding. We definiëren *shortcut learning* als het gebruik maken van gemakkelijk detecteerbare kenmerken die niet de volledige informatie van de taak representeren. We introduceren het *framework for synthetic shortcuts in vision-language models* om het gebruik van synthetische shortcuts tijdens training en evaluatie te analyseren, en laten zien dat contrastieve *vision-language* methoden vaak afhankelijk zijn van shortcuts, waardoor taakrelevante informatie wordt verwaarloosd. De studie benadrukt de noodzaak om *shortcut learning* te beperken om de robuustheid van *vision-language* representatieleren te verbeteren.

In Hoofdstuk 5 richten we ons op het verbeteren van de evaluatie van *vision-language*

modellen op de *image-text retrieval* taak in de context van conceptgranulariteit. We analyseren de conceptgranulariteit van bestaande *benchmarks* en stellen een nieuw evaluatiekader voor dat een taxonomie van verstoringen en een cross-modale evaluatiemetriek omvat. We evalueren vier *state-of-the-art vision-language* modellen op de benchmark, onderzoeken de impact van conceptgranulariteit op prestaties, en pleiten voor een verfijnde *evaluation pipeline* die beter de complexiteit van de realiteit weerspiegelt.

In Hoofdstuk 6 pakken we het probleem aan van het voorspellen van koopintentie in *cross-device* scenario's, waarbij we onderscheid maken tussen anonieme en geïdentificeerde sessies. We analyseren sessielogs van een Europees e-commerce platform om koopintentiesignalen te identificeren en ontwikkelen voorspellende modellen die sessiebaseerde en historische kenmerken in overweging nemen. We tonen aan dat koopintentie vroeg in de sessie kan worden voorspeld, waarbij gebruikers vaak overschakelen naar apparaten met grote schermen voor de definitieve aankopen. Dit werk draagt bij aan het begrip van gebruikersgedrag met meerdere apparaten voor zowel anonieme als geïdentificeerde sessies.

Tot slot, in Hoofdstuk 7 stellen we de *category-to-image retrieval* taak voor en onderzoeken we de impact van multimodale productrepresentaties op deze taak. We combineren tekstuele, visuele en attribuutinformatie en onderzoeken hun impact op prestaties in de context van categorieën met verschillende granulariteit. Experimenten tonen aan dat multimodale representaties over het algemeen de prestaties verbeteren, hoewel modellen die alleen afbeeldingen gebruiken multimodale modellen kunnen overtreffen voor algemene categorieën. De bevindingen benadrukken het belang van het overwegen van de wisselwerking tussen verschillende modaliteiten bij het bouwen van *retrieval models*.