Synthesis Lectures on Computer Vision

SYNTHESIS
COLLECTION OF TECHNOLOGY

Man Luo · Tejas Gokhale ·
Neeraj Varshney · Yezhou Yang ·
Chitta Baral

Advances in Multimodal Information Retrieval and Generation

Springer

Synthesis Lectures on Computer Vision

This series publishes on topics pertaining to computer vision and pattern recognition. The scope follows the purview of premier computer science conferences, and includes the science of scene reconstruction, event detection, video tracking, object recognition, 3D pose estimation, learning, indexing, motion estimation, and image restoration. As a scientific discipline, computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner. As a technological discipline, computer vision seeks to apply its theories and models for the construction of computer vision systems, such as those in self-driving cars/navigation systems, medical image analysis, and industrial robots.

Man Luo · Tejas Gokhale · Neeraj Varshney · Yezhou Yang · Chitta Baral

Advances in Multimodal Information Retrieval and Generation

Springer

Man Luo
AI research Scientist
Multimodal Cognitive AI Team
Intel Research Lab
Santa Clara, CA, USA

Tejas Gokhale
Department of Computer Science
and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, USA

Neeraj Varshney
School of Computing, Informatics
and Decision Systems Engineering
Arizona State University
Tempe, USA

Yezhou Yang
School of Computing, Informatics
and Decision Systems Engineering
Arizona State University
Tempe, USA

Chitta Baral
School of Computing, Informatics
and Decision Systems Engineering
Arizona State University
Tempe, USA

ISSN 2153-1056 ISSN 2153-1064 (electronic)
Synthesis Lectures on Computer Vision
ISBN 978-3-031-57815-1 ISBN 978-3-031-57816-8 (eBook)
https://doi.org/10.1007/978-3-031-57816-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

1 Introduction

Writing systems emerged simultaneously and independently in many ancient civilizations across the globe, including Mesopotamia, Egypt, China, India, and Mesoamerica. The invention of writing and scripts has had a tremendous impact on the trajectory of human civilization. Archaeological findings suggest the existence of much older proto-writing. Upper Paleolithic cave paintings in Europe contain dots and "Y" shaped symbols alongside paintings of animals and are conjectured to indicate the lunar mating cycle of animals—these markings are 20,000 years old and predate any other known writing or proto-writing systems [1]. The Vinca symbols (in present-day Balkans) are untranslated symbols that have been dated to be as old as the 7th millennium BCE—these symbols have been conjectured to contain information about who owned property, numerical symbols, emblems denoting communal identity or religious objects [2]. The Sumerian epic poem "Enmerkar and the Lord of Arrata" from around 1800 BCE describes a poetic version of the story of the invention of clay tablet writing systems [3]:

> Because the messenger's mouth was heavy and he couldn't repeat (the message), the Lord of Kulaba patted some clay and put words on it, like a tablet. Until then, there had been no putting words on clay.

Information storage and retrieval as concepts date back to early collections of clay tablets at Ebla (2900 BC, in contemporary Syria) and Nineveh (in present-day northern Iraq, near Mosul) which consisted of clay tablets inscribed with cuneiform writings. The Nineveh tablets, assembled by Ashurbanipal, King of the Assyrians (668–631 BC), is a collection that went beyond materials relevant to running a kingdom or a state, and the Epic of Gilgamesh was part of that collection [4]. The content of the clay tablets was multimodal in the sense that besides having graphic text (pictograms), and maps, they had physical tokens representing numbers, and also such tokens placed on clay envelops to describe the content of the envelop

M. Luo et al., *Advances in Multimodal Information Retrieval and Generation*, Synthesis Lectures on Computer Vision, https://doi.org/10.1007/978-3-031-57816-8_1

in a meta-data sense. The clay tablets in Ashurbanipal's collection had meta-data information recording their number in a series, filing method, and identification stamps. Multiple copies of various works in that collection suggest its simultaneous use by multiple people.

The library of Alexandria (285–145 BC) is considered perhaps the first library that used an indexing mechanism where books were sorted by alphabetical order of the first letter of the author's name. In the modern era, the Dewey Decimal Classification (DDC) debuted in 1876 and is widely used in libraries around the world. In the twentieth century, as libraries began to digitize their collections, digital library indexing systems were developed. These systems allowed for keyword searches, numeric or symbol tagging, and the use of Boolean operators. They also have made library collections accessible from any location with an internet connection. In the 21st century, there's a move towards linked data and Semantic Web technologies for libraries, enabling a higher level of interoperability and data sharing among different information systems. From ancient scroll repositories to digital databases, the history of library indexing systems shows the ever-increasing complexity of organizing and efficiently accessing information.

In the digital era, the term "information retrieval" was coined by Mooers in 1950 [5] and popularized by Fairthorne [6]. Some of the key developments in the early years of information retrieval (IR) in the digital era include development of indexing languages and their evaluations (such as the Cranfield experiments—1967), the first interactive retrieval systems such as DIALOG and MEDLINE [7], the Boolean model [8] of searching during retrieval, the formulation of "most closely match" and ranked-output [9], and the initial studies on user behavior to ground the concepts of "information need" and "relevance". The invention of the World in 1989 and DARPA's major evaluation exercises through the Text Retrieval Conferences (TRECs) that started in 1992 [10] had a huge impact on the subsequent development in the field of IR. For example, Google's early use of automated link analysis [11] to measure the relative importance of webpages and automated approaches to recognize spam, won over alternative approaches used by its competitors.

The use of machine learning methods in IR was another key landmark in the research evolution of IR. The implementation of machine learning algorithms for text classification laid the groundwork for their subsequent utilization in document ranking. Google's reverse image search, and content-based image retrieval methods in general use computer vision techniques, which now are mostly based on neural machine learning methods [12]. IR using neural methods was influenced by the vector representation of text and documents and was motivated by the need to address semantic understanding. Subsequent IR methods aimed at combining the semantic understanding and vocabulary mismatch aspect of neural IR with traditional IR challenges such as rare terms and intents [13].

Question answering (QA) is a closely related notion to IR where the query is formulated as a question in natural language, the retrieved information is answered, and the corpora (concerning which the question is answered) is often confined [14, 15]. In visual question answering (VQA), the question is asked concerning visual objects [16], and in visual-linguistic QA (VLQA) [17] and multimodal QA [18] the question is asked concerning multimodal objects. QA, IR, and prompt-based language generation [19] are getting integrated as users now expect aggregated information from multiple documents in response to their queries. This now involves the generation of text as well as images, and transformer-based large language models (LLMs) are now the key technology used for this [20, 21].

In this book, our emphasis is on multimodal information retrieval, specifically concentrating on text and image data. The traditional unimodal systems, limited to a single type of data, often fall short of capturing the complexity and richness of human communication and experience. In contrast, multimodal retrieval systems leverage the complementary nature of different data types to provide more accurate, context-aware, and user-centric search results. Text can provide specific details and context that images alone cannot convey. Conversely, images can instantly show concepts that might take longer to explain in words. Therefore, multimodal retrieval has wider applications in real world. For instance, consider you've previously visited a memorable location in New York City and captured a photo filled with landmarks and people. If you can't recall the place's name later, a multimodal system allows you to query "where is this place" along with the photo for identification. Healthcare is another domain where multimodal retrieval can be invaluable. Imagine a diagnostic support system that analyzes patients' electronic health records, which contain a mix of textual data (like doctor's notes), visual data (such as X-ray or MRI images), and even auditory data (like heart or lung sounds). A multimodal retrieval system can integrate these diverse data types to assist medical professionals in diagnosing complex conditions more accurately and swiftly.

In this book, we use the word "retrieval" in a broader sense to include the process of outputting aggregated information concerning prompted search queries to current-day generative AI models. In the rest of this chapter, we give a brief overview of the various aspects related to this. In Chap. 2, we discuss transformer-driven models for language, vision, and multimodal inputs; as transformers are key components of current-day generative AI models. In Chap. 3, we present various multimodal retrieval methods in the traditional sense of retrieval. In Chap. 4, we present generative AI models that generate multimodal content. In Chap. 5, we present how traditional retrieval can be used to augment generative models so that the resulting output is up-to-date and non-hallucinating.

1.1 Transformer-Driven Models for Language, Vision, and Multimodal Learning

This chapter focuses on the pivotal domains of artificial intelligence: language and vision, tracing their transformation through deep neural networks. The inception of convolutional neural networks (CNNs) revolutionized computer vision [22], while recurrent neural networks (RNNs) [23] marked a significant leap in natural language processing (NLP). Initially, in the early 2000s, the impact of neural networks was constrained by limited computational power and the scarcity of large-scale annotated datasets. However, this changed dramatically in 2016. Advancements in hardware enabled the training of complex models like ResNET, and the availability of datasets like ImageNet [24] showcased the true potential of deep neural networks. These advancements brought practical applications into everyday life, such as facial recognition technologies enhancing daily human interactions.

The year 2017 was a landmark in natural language processing with the introduction of the Transformer model [25]. Characterized by its self-attention, multi-head attention, and cross-attention mechanisms, the Transformer fundamentally altered the landscape of NLP. Following this, in 2018, the BERT model [26] emerged as the first language model based on the Transformer architecture, propelling NLP research to new heights. The success of BERT was underpinned by its self-supervised learning approach and deep contextual understanding of language. At its time, BERT was considered a large model, but in the context of 2023, it pales in comparison to state-of-the-art models like ChatGPT [27], which boast upwards of 175 billion parameters.

Transformers have since become the cornerstone of most cutting-edge models, not just in NLP but also in computer vision. This transition to vision was marked by the advent of the Vision Transformer (ViT) in 2020 [28]. The next frontier in AI research and application development is multimodal models. Given the multimodal nature of our world, understanding and integrating multiple forms of data is crucial for a more comprehensive understanding of our environment.

In this chapter, we will explore the evolution of both language and vision models, from their early development in the nascent years of deep learning to the contemporary era dominated by Transformer-based models. We will also discuss the most influential multimodal models that have laid the groundwork for many of the developments discussed throughout this book.

1.2 Multimodal Information Retrieval

In today’s digital world, where data presents itself in myriad forms—be it text, images, videos, or a combination of these—there’s an increasing need for systems that can effectively and efficiently retrieve the desired information. Multimodal Information Retrieval (MMIR) is a field that tackles the challenge of accurately retrieving specific information from a complex array of data types, including text, images, and videos, by developing systems that can efficiently search across these varied formats. In this chapter, we’ll take a comprehensive journey through the landscape of MMIR, especially focusing on its applications in text-image settings.

Initially, this chapter will outline the concepts of multimodal data and multimodal representation learning. Next, we will illuminate four key elements of Information Retrieval (IR), detailing their definitions and forms. Then we will categorize retrieval methods into two main approaches: text retrieval and multimodal retrieval. While text retrieval remains prevalent, our focus here leans more towards the dynamic field of multimodal retrieval, especially multimodal-queries retrieval, where queries seamlessly integrate components like text and images. In such systems, combining image and text information is crucial to comprehend queries and retrieve relevant documents. We will then discuss advanced multimodal transformer-based models, which transcend basic language and vision transformers. This includes a deep dive into the most exemplary models for handling multimodal queries. Following the exploration of MMIR methods, the chapter addresses their importance in key downstream applications such as question-answering and enhancing dialogue systems. Subsequent sections will investigate the evaluation metrics for IR systems, ranging from traditional metrics like precision and recall to more sophisticated measures. Finally, the chapter concludes by discussing the broader impacts of MMIR.

1.3 Multimodal Content Generation

Humans, since ancient times have observed the universe and tried to replicate it visually—in doing so, we have developed methods to create visual content. For example, cave paintings of hand prints or scenes depicting collaborative hunting tell us a story of a human community living together thousands of years ago. These images have allowed our ancestors to communicate what they saw- the environment, other creatures, other humans, and their interactions with them. Content creation, storage, and dissemination are thus an integral part of the history of our civilization.

In Chap. 4, we will learn about the research area of content generation, with special emphasis on vision-language content generation. This chapter sets up fundamental concepts in this domain such as conditional generative models and discusses several modeling techniques that use generative adversarial networks or diffusion models. We will also set up a taxonomy for conditional image generation which includes categorical conditions (e.g. using class labels as inputs to content generation models), visual conditions (such as sketches or semantic label maps), and the recent explosion of text-to-image generation (generating images directly from natural language descriptions). We will discuss text-to-image (T2I) generation in detail, by focusing on the two dominating models for T2I: GANs and diffusion models. We will learn how recent developments have resulted in many applications of T2I models in image editing, compositional generation, and iterative generation. We will also discuss several applications of text-guided generative models, for instance in generating audio, video, three-dimensional structures and assets, and other applications. We will also discuss the task of image and video captioning.

Over the last decade, sophisticated modeling strategies have emerged for image generation, language generation, audio generation, and many other forms of content generation. The models have leveraged the availability of web-scale datasets to develop training protocols that have resulted in highly realistic content generation. This is quickly creating a new wave in the digital media industry. This excitement in both academic and industrial circles has also been accompanied by challenges related to the robustness, reliability, and risks of using content-generation models. We will discuss some of the recent efforts of developing evaluation strategies and benchmarks that could potentially address some of the challenges by providing quantifiable and grounder insights about the capabilities of content generation models and their failure modes to better inform users of these technologies.

1.4 Retrieval Augmented Modeling

The previous chapters cover the motivations behind information retrieval, its fundamental principles, core components, and the various strategies employed to achieve effective retrieval. These include multimodal retrieval and generative retrieval, each with its motivations for study. Moving forward, we'll place a special emphasis on the integration of retrieval techniques with language models, a concept known as retrieval-augmented modeling.

The motivation behind studying retrieval augmented modeling is to create language models that not only understand the given input but also tap into external knowledge sources for more comprehensive and precise responses. In Sect. 5.1, we'll elaborate on the diverse ways in which language models can harness retrieved information to enhance their responses. This encompasses enriching the input to provide context, refining intermediate layers to improve comprehension, and augmenting the output for more informed responses.

The crucial aspect of retrieval-augmented modeling lies in the training of both the retriever and language models. In Sect. 5.2, we'll discuss three distinct strategies for training these models: independent training, sequential training, and joint training. In Sect. 5.3, we'll outline the various types of information that can be harnessed, such as knowledge, similar examples, and generated context, to produce informed responses. The significance here is to understand the diverse sources of information that can contribute to more accurate outputs from the model.

Shifting our focus to practical applications, Sect. 5.4 will examine the real-world impact of retrieval augmented language models. We'll explore their applications, including fact-checking and addressing the issue of factual 'hallucinations' that sometimes occur with large language models. Finally, in Sect. 5.5, we will shift our focus to leveraging the generation ability of large language models to improve retrieval performance.

1.5 Target Audience

This book covers inter-disciplinary topics, spanning information retrieval, computer vision, natural language processing, machine learning, and others. The book is intended to be a resource for advanced undergraduates, graduate students, faculty, and researchers working in these fields, adjacent areas, or those seeking an introduction to frontier research in this area. We intend to make this book accessible to readers from all of these communities to foster active dialog and exchange of ideas. Frontiers of academic research in this domain are closely connected with potential applications, such as search engines, chat-bots, AI assistants, etc. This makes the book a resource for practitioners, engineers, and designers working towards the development of such products.

In addition to this book, we have also been involved in building a community of researchers interested in adjacent topics. We were invited to organize a workshop on multimodal information retrieval at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) in June 2022. This workshop, titled "Open Domain Retrieval under Multimodal Settings" (or ODRUM for short) was designed to bring together prominent researchers from multiple research fields and perspectives such as information retrieval, natural language processing (NLP), computer vision (CV), and knowledge representation and reasoning (KRR). This workshop aimed to address the relatively nascent direction of information retrieval with queries that may come from multiple modalities (such as text, images, videos, audio, etc.), or multiple formats (paragraphs, tables, charts, etc.). The reader is encouraged to avail of the publicly available video recordings, slides, accepted papers, additional reading materials, and discussion directions that could spark open research questions in multimodal information retrieval. More details can be found at the workshop website https://asu-apg.github.io/odrum/archive_2022.html.

References

1. Bennett Bacon, Azadeh Khatiri, James Palmer, Tony Freeth, Paul Pettitt, and Robert Kentridge. An upper palaeolithic proto-writing system and phenological calendar. *Cambridge Archaeological Journal*, 33(3):371389, 2023. https://doi.org/10.1017/S0959774322000415.
2. Sarunas Milisauskas and Janusz Kruk. Middle neolithic/early copper age, continuity, diversity, and greater complexity, 5500/5000–3500 bc. *European Prehistory: A Survey*, pages 223–291, 2011.
3. Peter T Daniels. The study of writing systems. *The world's writing systems*, pages 3–17, 1996.
4. Irving Finkel. Assurbanipal's library. *Libraries before Alexandria: Ancient Near Eastern Traditions*, page 367, 2019.
5. Calvin Mooers. Information retrieval viewed as temporal signaling. In *Proceedings of the international congress of mathematicians*, volume 1, pages 572–573, 1950.
6. RA Fairthorne. Towards information retrieval. *Journal of the Operational Research Society*, 14(2):215–216, 1963.

7. R Brian Haynes, Nancy Wilczynski, K Ann McKibbon, Cynthia J Walker, and John C Sinclair. Developing optimal search strategies for detecting clinically sound studies in medline. *Journal of the American Medical Informatics Association*, 1(6):447–458, 1994.
8. Frederick Wilfrid Lancaster and Emily Gallup. Information retrieval on-line. Technical report, 1973.
9. Stephen E Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.
10. DK Harman. Overview of the first text retrieval conference (trec-1). *NIST Special Publication*, pages 500–207, 1992.
11. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bring order to the web. Technical report, Technical report, stanford University, 1998.
12. Ricardo da Silva Torres and Alexandre X Falcao. Content-based image retrieval: theory and applications. *RITA*, 13(2):161–185, 2006.
13. Bhaskar Mitra, Nick Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.
14. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 201. https://doi.org/10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.
15. Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, 2017a. Association for Computational Lingu. https://doi.org/10.18653/v1/P17-1171. URL https://aclanthology.org/P17-1171.
16. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015. https://doi.org/10.1109/ICCV.2015.279. URL https://doi.org/10.1109/ICCV.2015.279.
17. Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. Visuo-linguistic question answering (vlqa) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4606–4616, 2020.
18. Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504, 2022a.
19. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
20. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017a. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
21. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,