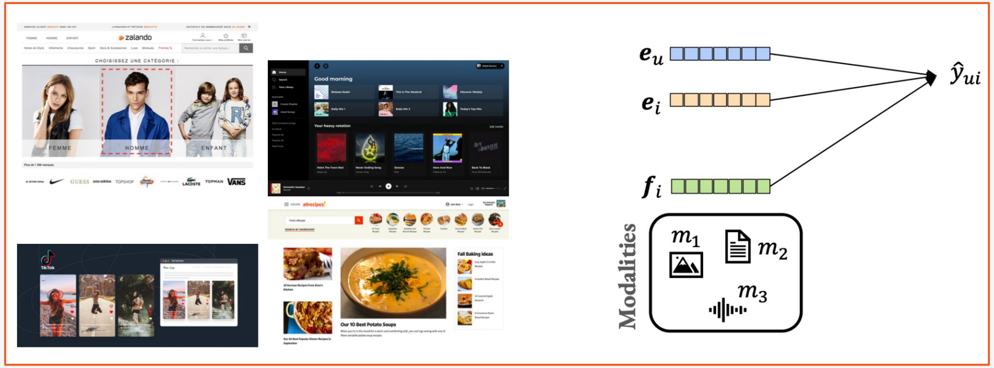
06

Multimodal Feature Extraction in Recommendation

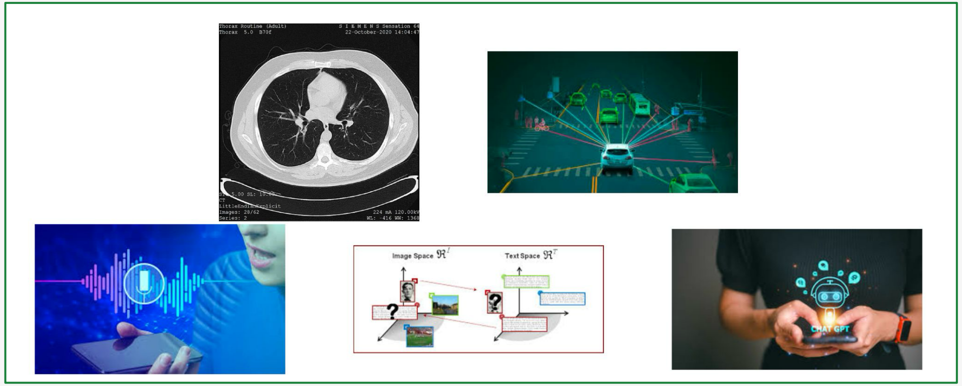
Multimedia recommendation

Multimedia recommender systems [34, 35] augment items' representation through their multimodal features extracted from images, texts, or audio tracks describing them.



[34] Deldjoo et al., Recommender Systems Leveraging Multimedia Content. ACM Comput. Surv. 53(5): 106:1-106:38 (2021)
[35] Malitesta et al., Formalizing Multimedia Recommendation through Multimodal Deep Learning. Trans. Recomm. Syst. 3(3): 37:1-37:33 (2025)

Multimodal deep learning

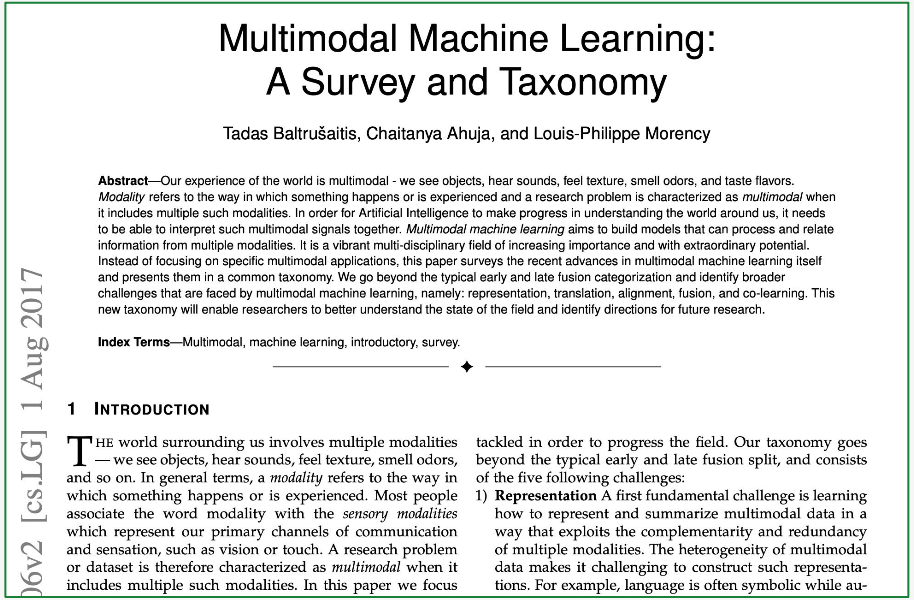


Applications of multimodal DL: medical imaging [36], autonomous driving [37], speech/emotion recognition [38], multimedia retrieval [39], multimodal large language modelling [40].

[36] Tang et al., IEEE TIP 2022, MATR: multimodal medical ...
[37] Caesar et al., CVPR 2020, nuscenes: A multimodal dataset ...
[38] Pan et al., ACL 2022, Leveraging unimodal self-supervised learning ...
[39] Li et al., SIGIR 2021, Hybrid fusion with intra- and cross-modality ...
[40] Yin et al., CoRR 2023, A survey on multimodal large language models.

Multimodal DL pipeline

Some works have tried to outline, categorize, and formalize the core concepts behind multimodality in deep learning [39] through a pipeline: representation, translation, alignment, fusion, and co-learning.



Multimodal Machine Learning:
A Survey and Taxonomy

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency

Abstract—Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors. *Modality* refers to the way in which something happens or is experienced and a research problem is characterized as *multimodal* when it includes multiple such modalities. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. *Multimodal machine learning* aims to build models that can process and relate information from multiple modalities. It is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential. Instead of focusing on specific multimodal applications, this paper surveys the recent advances in multimodal machine learning itself and presents them in a common taxonomy. We go beyond the typical early and late fusion categorization and identify broader challenges that are faced by multimodal machine learning, namely: representation, translation, alignment, fusion, and co-learning. This new taxonomy will enable researchers to better understand the state of the field and identify directions for future research.

Index Terms—Multimodal, machine learning, introductory, survey.

06v2 [cs.LG] 1 Aug 2017

1 INTRODUCTION

THE world surrounding us involves multiple modalities — we see objects, hear sounds, feel texture, smell odors, and so on. In general terms, a *modality* refers to the way in which something happens or is experienced. Most people associate the word modality with the *sensory modalities* which represent our primary channels of communication and sensation, such as vision or touch. A research problem or dataset is therefore characterized as *multimodal* when it includes multiple such modalities. In this paper we focus

tackled in order to progress the field. Our taxonomy goes beyond the typical early and late fusion split, and consists of the five following challenges:

1) **Representation** A first fundamental challenge is learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representa- tions. For example, language is often symbolic while au-

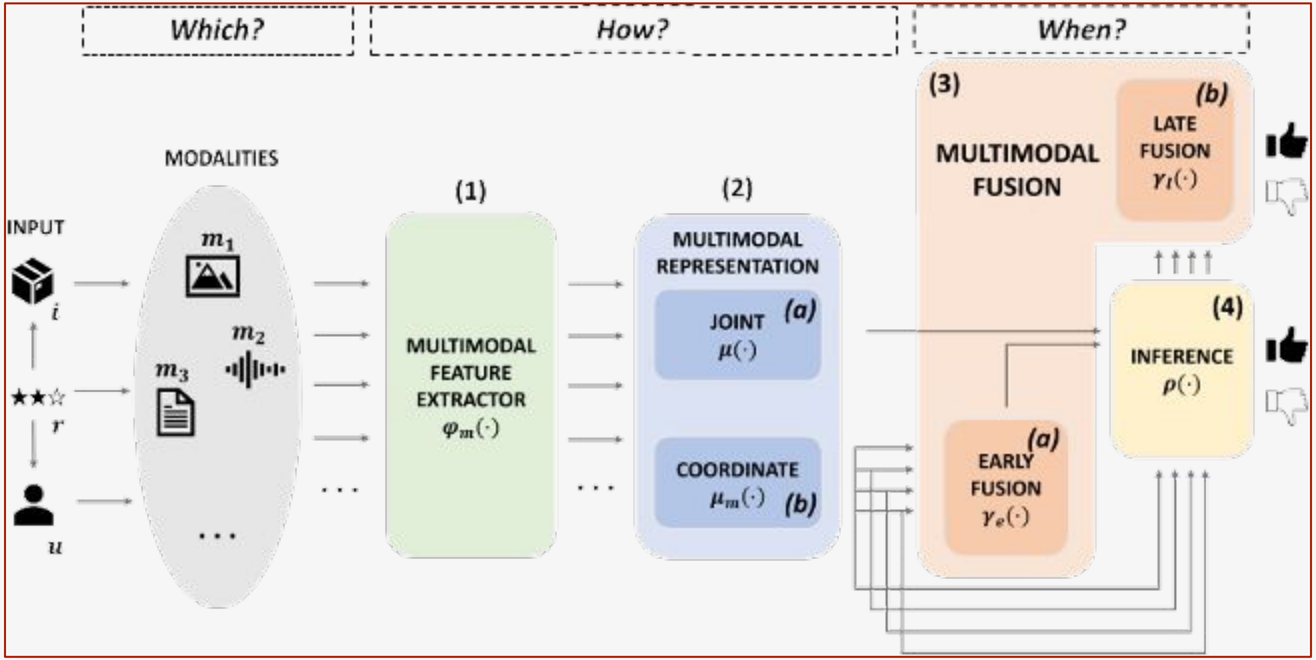
[41] Baltrusaitis et al., Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41(2): 423-443 (2019).

Bridging multimodal DL and RSs

However, recommendation lacks a shared theoretical and applicative formalization to align the multimedia recommendation problem with the same formal pipeline proposed in multimodal deep learning.

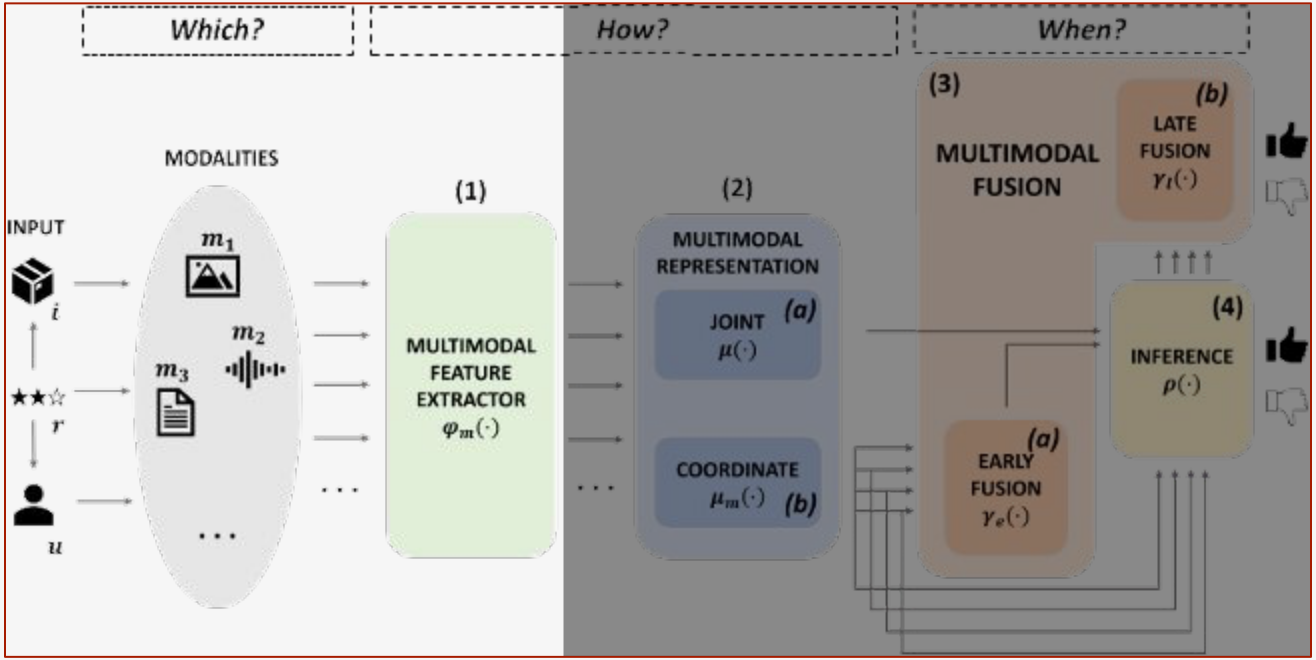
[35] Malitesta et al., Formalizing Multimedia Recommendation through Multimodal Deep Learning. Trans. Recomm. Syst. 3(3): 37:1-37:33 (2025)

The multimodal pipeline for RSs



[35] Malitesta et al., Formalizing Multimedia Recommendation through Multimodal Deep Learning. Trans. Recomm. Syst. 3(3): 37:1-37:33 (2025)

Our main focus for this tutorial



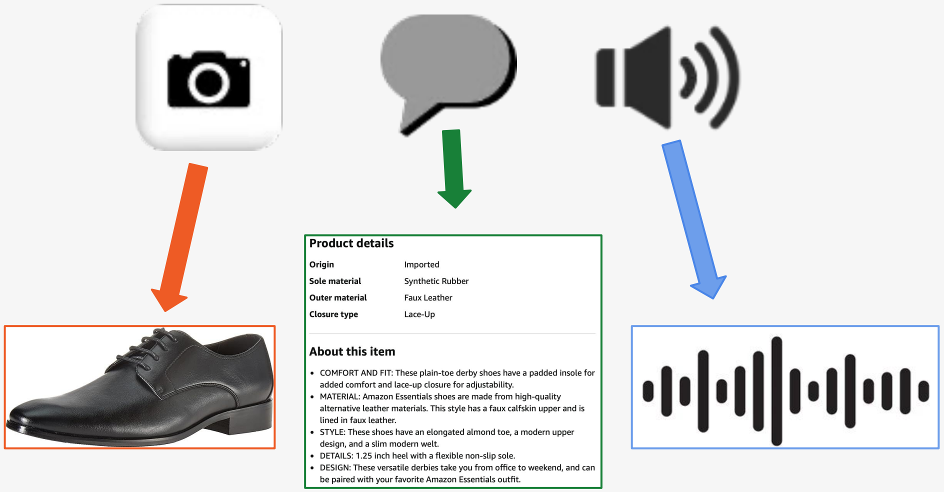
[35] Malitesta et al., Formalizing Multimedia Recommendation through Multimodal Deep Learning. Trans. Recomm. Syst. 3(3): 37:1-37:33 (2025)

Multimodal data input

Formally, we define $m \in M$ as an admissible modality for the system (i.e., $M = \{v, t, a\}$).

Let $x \in X$ be an input to the recommender system, whose set of available modalities is indicated as $M_x \subseteq M$.

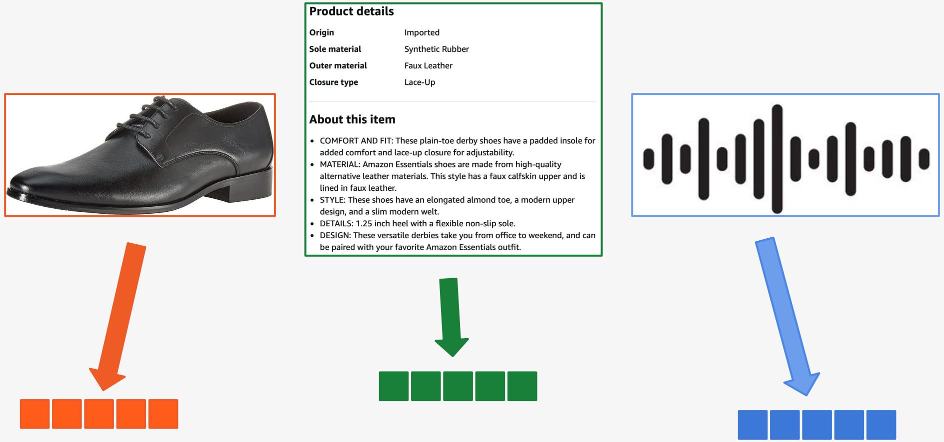
Multimodal data input



We represent the content data of input x in modality m as $c_x^{(m)}$, with $m \in M_x$, and the vector of content data for input x in all modalities as c_x.

Multimodal feature processing

Our schema introduces a Feature Extractor (FE). Let $c_x^{(m)}$ be the content data for input x in modality. $m \in Mx$. Then, let $\varphi_m(\cdot)$ be the feature extractor function for the modality m.



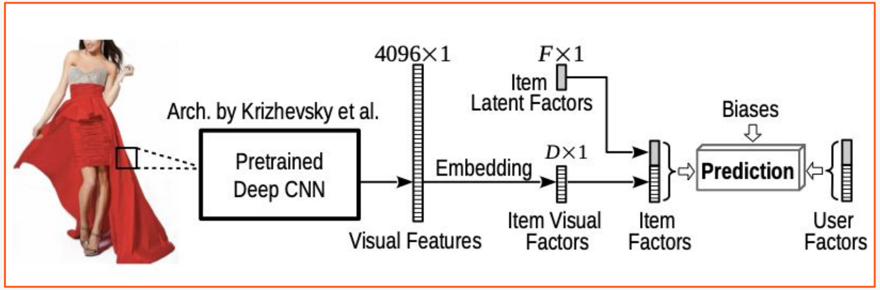
Multimodal feature processing

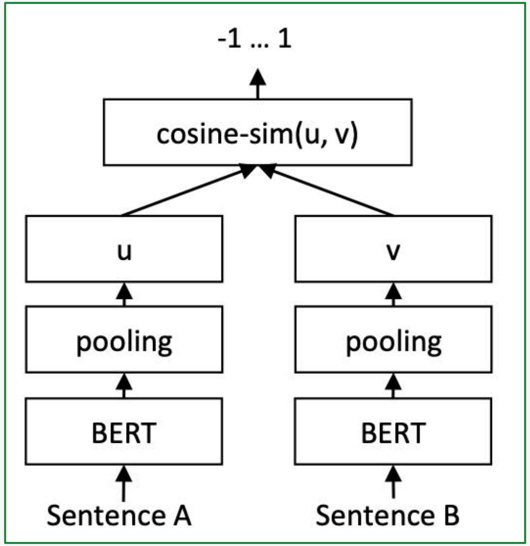
We define the feature extraction process in the modality *m* as:

$$\overline{c}_x^{(m)} = \varphi_m(c_x^{(m)}) \quad \forall m \in \mathcal{M}_x,$$

where $c_x^{(m)}$ is the extracted feature for input *x* in modality m. We use the notation $\overline{\mathrm{c}}_x = [\overline{c}_x^{(0)}, \overline{c}_x^{(1)}, \ldots, \overline{c}_x^{(|\mathcal{M}_x|-1)}]$ to refer to the vector of extracted features for input x in all modalities.

Multimodal feature processing





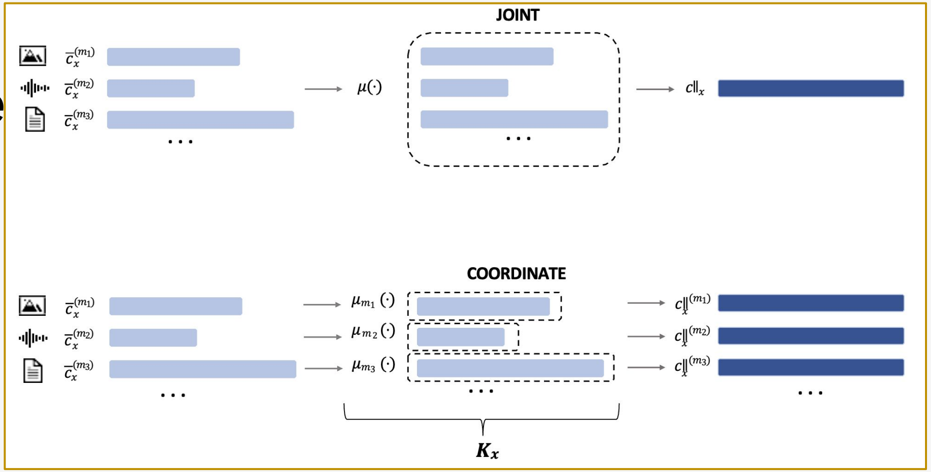
[42] He and McAuley: VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. AAAI 2016: 144-150
[43] Reimers and Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP/IJCNLP (1) 2019: 3980-3990

Multimodal feature processing

The next step is to design a Representation strategy to handle the relationships among modalities and eventually inject such data into the recommender system.

Multimodal feature processing

The literature follows two main approaches: *Joint* and *Coordinate*. Whateve the chosen strategy, the final multimodal representation is indicated as $\tilde{c}_x$.



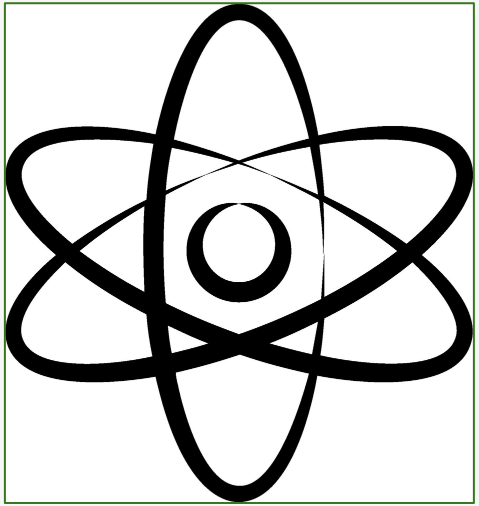
Multimodal feature fusion

Early: first we fuse...

$$\tilde{c}_x = \gamma_e(\tilde{\mathrm{c}}_x).$$

... then, we predict

$$\hat{y} = \rho(\tilde{c}_x).$$



Multimodal feature fusion

Late: first we predict...

$$\hat{y}^{(m)} = \rho(\tilde{c}_x^{(m)}) \quad \forall m \in \mathcal{M}_x.$$

... then, we fuse

$$\hat{y} = \gamma_l(\widehat{\mathrm{y}}).$$