Multimodal Data Fusion

Subjects: Computer Science, Information Systems

Contributor: Maciej Pawłowski, Anna Wróblewska, Sylwia Sysko-Romańczuk

Data processing in robotics is currently challenged by the effective building of multimodal and common representations. Tremendous volumes of raw data are available and their smart management is the core concept of multimodal learning in a new paradigm for data fusion.

Keywords: multimodal representation ; multimodal learning ; data fusion ; deep learning in sensor systems

1. Introduction

The emergence of our hyper-connected and hyper-digitalized world (IoT, ubiquitous sensing, etc.) requires any education organization to have the ability to handle a system that produces huge amounts of different data. A key area of research in multimodal data is the process of building multimodal representations, the quality of which determines the modeling and prediction of organizational learning.

Multimodal learning involves working with data that contain multiple modalities, such as text, images, audio, and numerical or behavioral data. The interest in multimodal learning began in the 1980s and has gained popularity since then. In one of the first papers on the subject [1], the authors demonstrated that acoustic and visual data could be successfully combined in a speech recognition sensor system. By integrating both of the modalities they outperformed the audio data model.

Further research has shown that modalities can be complementary and a carefully chosen data fusion leads to significant model performance improvements [2][3][4]. Furthermore, modalities might naturally coexist, and the analyzed machine learning (ML) problem cannot be solved unless all modalities are considered [5]. Finally, recent improvements in multisensor solutions [6] have offered high-quality modalities, which require an appropriate fusion technique to fit to the ML problem.

Choosing, or even designing, the right data fusion technique is fundamental to determining the quality and performance of data fusion strategies to be undertaken by data scientist teams in organizations. Moreover, data fusion strategies are the central concept of multimodal learning. This concept could be applicable to numerous machine learning tasks with all kinds of modalities (a so-called universal technique). According to [4], a universal approach has not been established yet. The research work conducted by the authors shows that all current state-of-the-art data fusion models suffer from two main issues. Either they are task-specific or too complicated or they lack interpretability and flexibility.

2. Types of Modalities

This section introduces the types of modalities that can be encountered while working on any machine learning problem. According to [7], the researchers identify four main groups of modalities:

- Tabular data: observations are stored as rows and their features as columns;
- Graphs: observations are vertices and their features are in the form of edges between individual vertices;
- Signals: observations are files of appropriate extension (images:.jpeg, audio:.wav, etc.) and their features are the numerical data provided within files;
- Sequences: observations are in the form of characters/words/documents, where the type of character/word corresponds to features.

The research includes modalities coming from each of the identified groups: labels are an example of tabular data; reviews, titles, and descriptions represent textual data (sequences); images of products and movie posters are examples of visual data;, and relations between movies (movies seen by one user) are graphs. **Table 1** introduces some other examples of works on various combinations of modalities. Additionally, in [8], the authors concentrated on the image segmentation task to learn the optimal joint representation from rich and complementary features of the same scene based on images retrieved from different sensors. All these studies showed no restrictions on modalities that can be studied together to solve a specific problem. Moreover, they proved that the combination of modalities boosts performance; the multimodal model achieved better results than the unimodal models.

Table 1. Examples of multimodal tasks.

Modalities	Article	Overview
Images, time series, and tabular	[9]	Prediction of Alzheimer’s disease based on magnetic resonance imaging and positron emission tomography (images) that are performed multiple times on one patient within specified periods of time (time series). Patient demographics and genetic data are also taken into account (tabular).
Audio, video, and event streams	[10]	Behavioral analysis and emotion and stress prediction. Analyzed data consist of 45-min recordings of students during the final exam period. They are recorded with the use of cameras (video), thermal physiological measurements of the heart, breathing rates (event streams), and lapel microphones (audio).
Text and images	[11]	Question answering based on images containing some textual data.
Images, text, and graphs	[12][13][14]	Outfit/movie recommender systems. Movies are recommended based on plot (text), poster (image), liked and disliked movies, and cast (graphs). Outfits are chosen based on product features in images and text descriptions.

Furthermore, these works encapsulate all multimodal fusion techniques that the researchers examine in the research: early fusion, late fusion, and sketch. They are proven effective but have not been compared yet.

3. Multimodal Representation

Learning to represent is an unsupervised task, and there is no single way to describe a good representation. However, several works have identified the main features demanded while deriving any numerical representation of a given modality.

The problem of unimodal representation has already been solved with modality-dedicated models, such as BERT [15] for textual data, ResNet [16] for images, etc. However, a universal method that could be applied to any machine learning task when it comes to multimodal data has not been established [4].

Bengio et al. [17] characterized several features that an appropriate vector representation should possess, including:

- Smoothness: A transformation should preserve objects similarities, expressed mathematically as $x \approx y \Rightarrow f(x) \approx f(y)$. For instance, the words “book” and “read” are expected to have similar embeddings;
- Manifolds: probability mass is concentrated within regions of lower dimensionality than the original space, e.g., we can expect the words “Poland”, “USA”, and “France” to have embeddings within a certain region, and the words “jump”, “ride”, and “run” in another distinct region;
- Natural clustering: categorical values could be assigned to observations within the same manifold, e.g., a region with the words “Poland”, “USA”, and “France” can be described as “countries”;
- Sparsity: given an observation, only a subset of its numerical representation features should be relevant. Otherwise, we end up with complicated embeddings whose highly correlated features may lead to numerous ambiguities.

For multimodal representation, ref. [18] identifies more factors that should be taken into account: (1) the similarity between individual modalities should be preserved in their joint representation and (2) robustness to the absence of some modalities; it should still be possible to create multimodal embedding.

4. Multimodal Data Fusion

Multimodal data fusion is an approach for combining single modalities to derive multimodal representation. A few issues should be taken into account [4] when it comes to fusing several modalities:

- Intermodality: the combination of several modalities, which leads to better and more robust model predictions [2];

- Cross-modality: this aspect assumes inseparable interactions between modalities. No reasonable conclusions can be drawn from data unless all their modalities are joined [19];

- Missing data: for some cases, particular modalities might not be available. An ideal multimodal data fusion algorithm is robust to missing modalities and uses others to compensate for the widespread information loss in recommender systems.

Classically, the existing multimodal representation techniques are divided into two categories [2][20]: early (feature) and late (decision) fusion. In the early feature approach, all modalities are combined. This is usually achieved by concatenating their vector representations at an initial stage, and then one model is trained [21]. In the case of late fusion, several independent models concerning each modality are trained, then their outputs are connected. The connection can be made arbitrarily. One can average the outputs and pick the most frequent one (in classification tasks), or concatenate them and build a model to obtain a final output [21]. Neither of these data fusion approaches can be described as the best one [20]; both have been proven to yield promising results in various scenarios.

4.1. Deep Learning Models

The most popular multimodal fusion techniques are based on deep learning solutions. The authors of [4] describe such architecture ideas, along with their most representative cases. Four prominent approaches are deep belief nets, stacked autoencoders, convolution networks, and recurrent networks. However, despite their promising results in the field of multimodal data fusion, deep learning models suffer from two main issues [4]. Firstly, deep learning models contain enormous free weights, especially parameters associated with a modality that brings little information. This results in high resource requirements; an undesirable feature in a production scenario. Secondly, multimodal data usually come from very dynamic environments. Therefore, there is a need for a flexible model that can quickly adapt to all changes in data.

Enormous computational requirements and low flexibility suggest exploring other techniques applied to any task, despite the types of modalities. Furthermore, the authors of [4] suggest that these ideas can be combined with deep learning techniques and existing multimodal models to obtain state-of-the-art solutions, which would be applicable in every field and robust to all data imperfections (missing modalities, data distribution changes over time in a production case, etc.). According to [22], the best approach is based on deep learning; the challenge is modality fusion. One of the possible solutions is the use of hashing methods. The following section discusses the strengths and weaknesses of such algorithms.

4.2. Hashing Ideas

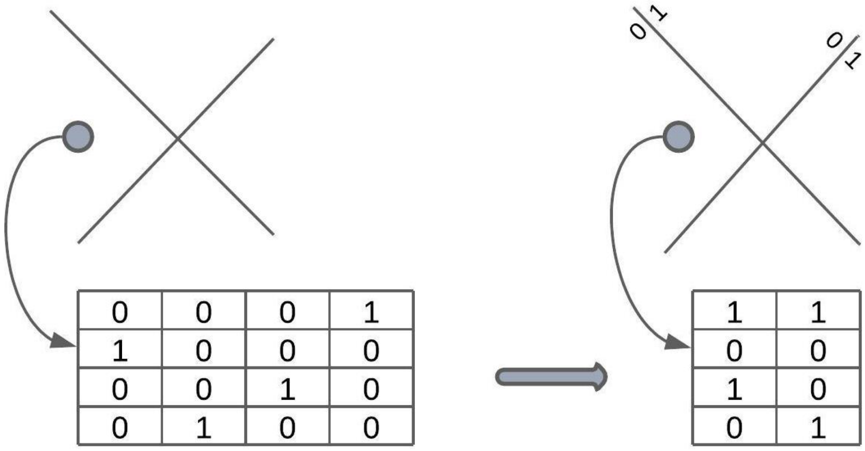
Another promising approach in multimodal data fusion is associated with hashing models. They identify manifolds in the original space and then transform data to lower-dimensional spaces while preserving observation similarities. Such algorithms can construct multimodal representation on the fly and have been proven effective in information retrieval problems [3], recommendation systems [23], and object detection cases. The main advantages [3] of hashing methods are that they (1) are cost-effective in terms of memory usage, (2) detect and work within manifolds, (3) preserve semantic similarities between points, (4) are usually data-independent, and (5) are suitable for production cases as they are robust to any data changes.

Unfortunately, hashing methods struggle with one issue. The mapping of high dimensional data into much simpler representations can result in the loss of certain information about specific observations [3]. Therefore, it has to be verified if hashing ideas can be applied to other fields apart from similarity search tasks. Perhaps their ability to combine multiple modalities while maintaining low costs and robustness to data changes recompenses the lost information.

4.3. Sketch Representation

The sketch representation has already been proven effective if fed with visual, behavioral, and textual data for the recommendation and similarity search tasks [23]. The idea of this representation comes from combining two algorithms: locality sensitive hashing and count-min sketch. All modalities are transformed into mutual space with the use of hash functions. Generally, a sketch is a one-hot sparse matrix containing all combined modalities. Hash functions make the representation modality independent, robust to missing modalities, and easily interpreted. Furthermore, modalities can be added to the sketch on the fly, which is extremely important in a production scenario.

In the research, the researchers slightly modify this sketch representation to the binarized form, see **Figure 1**. Instead of representing an observation with a subspace ID, it can be represented as a set of binary features. Then, 0 and 1 represent where the point lies concerning a single hyperplane. Such a sketch should preserve more information about a single observation.



0	0	0	1
1	0	0	0
0	0	1	0
0	1	0	0

1	1
0	0
1	0
0	1

Figure 1. The idea of binarizing the sketch. Instead of representing an observation with a subspace ID, it can be represented as a set of binary features. Then, 0 and 1 represent where the point lies concerning a single hyperplane. Such a sketch consumes much less memory and perhaps preserves more information about a single observation.

5. Multimodal Model Evaluation

Evaluating the multimodal data fusion algorithm is not straightforward, and no universal metric would measure the aspect of captured inter- and cross-modalities [19]. However, we can assess whether learning from multiple data types simultaneously enhances task performance.

The most popular way of verifying the quality of the multimodal fusion model [20] is to compare its performance scores (precision, AUC, etc.) to those achieved by models considering single modalities. With such an approach, we can state whether and to what extent combining modalities brings new information. The researchers also aim to preserve all similarities between observations, i.e., similar observations should be comparable in their multimodal representations. Therefore, several works [21][23] have compared their multimodal models to NN algorithms, which serve as a good baseline. Lastly, multimodal models should be tested when adjusting additional modalities. In certain cases [21], adding new modalities slightly improves the results while the training time increases dramatically. As a result, the model might be unfeasible in production scenarios despite its excellent performance. Therefore, we should not only focus on the scores the model achieves but also consider its flexibility and simplicity.

References

1. Yuhas, B.; Goldstein, M.; Sejnowski, T. Integration of acoustic and visual speech signals using neural networks. IEEE Commun. Mag. 1989, 27, 65–71.
2. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 41, 423–443.
3. Cao, W.; Feng, W.; Lin, Q.; Cao, G.; He, Z. A Review of Hashing Methods for Multimodal Retrieval. IEEE Access 2020, 8, 15377–15391.

4. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A Survey on Deep Learning for Multimodal Data Fusion. Neural Comput. 2020, 32, 829–864.
5. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf. Fusion 2023, 91, 424–444.
6. Tsanousa, A.; Bektsis, E.; Kyriakopoulos, C.; González, A.G.; Leturiondo, U.; Gialampoukidis, I.; Karakostas, A.; Vrochidis, S.; Kompatsiaris, I. A Review of Multisensor Data Fusion Solutions in Smart Manufacturing: Systems and Trends. Sensors 2022, 22, 1734.
7. Varshney, K. Trust in Machine Learning, Manning Publications, Shelter Island, Chapter 4 Data Sources and Biases, Section 4.1 Modalities. Available online: https://livebook.manning.com/book/trust-in-machine-learning/chapter-4/v-2/ (accessed on 23 March 2021).
8. Zhang, Y.; Sidibé, D.; Morel, O.; Mériaudeau, F. Deep multimodal fusion for semantic image segmentation: A survey. Image Vis. Comput. 2021, 105, 104042.
9. El-Sappagh, S.; Abuhmed, T.; Riazul Islam, S.; Kwak, K.S. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing 2020, 412, 197–215.
10. Jaiswal, M.; Bara, C.P.; Luo, Y.; Burzo, M.; Mihalcea, R.; Provost, E.M. MuSE: A Multimodal Dataset of Stressed Emotion. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Paris, France, 2020; pp. 1499–1510.
11. Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; Rohrbach, M. Towards VQA Models That Can Read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8309–8318.
12. Rychalska, B.; Basaj, D.B.; Dabrowski, J.; Daniluk, M. I know why you like this movie: Interpretable Efficient Multimodal Recommender. arXiv 2020, arXiv:2006.09979.
13. Laenen, K.; Moens, M.F. A Comparative Study of Outfit Recommendation Methods with a Focus on Attention-based Fusion. Inf. Process. Manag. 2020, 57, 102316.
14. Salah, A.; Truong, Q.T.; Lauw, H.W. Cornac: A Comparative Framework for Multimodal Recommender Systems. J. Mach. Learn. Res. 2020, 21, 1–5.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Toronto, AB, Canada, 2019; pp. 4171–4186.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
17. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 1798–1828.
18. Srivastava, N.; Salakhutdinov, R. Multimodal learning with deep Boltzmann machines. J. Mach. Learn. Res. 2014, 15, 2949–2980.
19. Frank, S.; Bugliarello, E.; Elliott, D. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. arXiv 2021, arXiv:2109.04448.
20. Gallo, I.; Calefati, A.; Nawaz, S. Multimodal Classification Fusion in Real-World Scenarios. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 36–41.
21. Kiela, D.; Grave, E.; Joulin, A.; Mikolov, T. Efficient Large-Scale Multi-Modal Classification. arXiv 2018, arXiv:1802.02892.
22. Bayoudh, K.; Knani, R.; Hamdaoui, F.; Mtibaa, A. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. Vis. Comput. 2022, 38, 2939–2970.
23. Dabrowski, J.; Rychalska, B.; Daniluk, M.; Basaj, D.; Goluchowski, K.; Babel, P.; Michalowski, A.; Jakubowski, A. An efficient manifold density estimator for all recommendation systems. arXiv 2020, arXiv:2006.01894.

Retrieved from https://encyclopedia.pub/entry/history/show/124708