Title: Multimodal representation learning

URL: https://en.wikipedia.org/wiki/Multimodal_representation_learning

PageID: 79726065

Categories: Category:Machine learning

Source: Wikipedia (CC BY-SA 4.0).

Multimodal representation learning is a subfield of representation learning focused on integrating and interpreting information from different modalities , such as text, images, audio, or video, by projecting them into a shared latent space. This allows for semantically similar content across modalities to be mapped to nearby points within that space, facilitating a unified understanding of diverse data types. [1] By automatically learning meaningful features from each modality and capturing their inter-modal relationships, multimodal representation learning enables a unified representation that enhances performance in cross-media analysis tasks such as video classification, event detection, and sentiment analysis. It also supports cross-modal retrieval and translation, including image captioning, video description, and text-to-image synthesis.

Motivation

The primary motivations for multimodal representation learning arise from the inherent nature of real-world data and the limitations of unimodal approaches. Since multimodal data offers complementary and supplementary information about an object or event from different perspectives, it is more informative than relying on a single modality. [1] A key motivation is to narrow the heterogeneity gap that exists between different modalities by projecting their features into a shared semantic subspace. This allows semantically similar content across modalities to be represented by similar vectors, facilitating the understanding of relationships and correlations between them. Multimodal representation learning aims to leverage the unique information provided by each modality to achieve a more comprehensive and accurate understanding of concepts.

These unified representations are crucial for improving performance in various cross-media analysis tasks such as video classification, event detection, and sentiment analysis. They also enable cross-modal retrieval, allowing users to search and retrieve content across different modalities. [2] Additionally, it facilitates cross-modal translation, where information can be converted from one modality to another, as seen in applications like image captioning and text-to-image synthesis. The abundance of ubiquitous multimodal data in real-world applications, including understudied areas like healthcare, finance, and human-computer interaction (HCI), further motivates the development of effective multimodal representation learning techniques. [3]

Approaches and methods

Canonical-correlation analysis based methods

Canonical-correlation analysis (CCA) was first introduced in 1936 by Harold Hotelling [4] and is a fundamental approach for multimodal learning. CCA aims to find linear relationships between two sets of variables. Given two data matrices X ∈ R n × p {\displaystyle X\in \mathbb {R} ^{n\times p}} and Y ∈ R n × q {\displaystyle Y\in \mathbb {R} ^{n\times q}} representing different modalities, CCA finds projection vectors w x ∈ R p {\displaystyle w_{x}\in \mathbb {R} ^{p}} and w y ∈ R q {\displaystyle w_{y}\in \mathbb {R} ^{q}} that maximizes the correlation between the projected variables:

ρ = max w x , w y w x ■ Σ x y w y w x ■ Σ x x w x w y ■ Σ y y w y {\displaystyle \rho =\max _{w_{x},w_{y}}{\frac {w_{x}^{\top }\Sigma _{xy}w_{y}}{{\sqrt {w_{x}^{\top }\Sigma _{xx}w_{x}}}{\sqrt {w_{y}^{\top }\Sigma _{yy}w_{y}}}}}}

such that Σ x x {\displaystyle \Sigma _{xx}} and Σ y y {\displaystyle \Sigma _{yy}} are the within-modality covariance matrices, and Σ x y {\displaystyle \Sigma _{xy}} is the between-modality covariance matrix. However, standard CCA is limited by its linearity, which led to the development of nonlinear extensions, such as kernel CCA and deep CCA.

Kernel CCA

Kernel canonical correlation analysis (KCCA) extends traditional CCA to capture nonlinear relationships between modalities by implicitly mapping the data into high dimensional feature spaces using kernel functions . Given kernel functions K x {\displaystyle K_{x}} and K y {\displaystyle K_{y}} with corresponding Gram matrices K x ∈ R n × n {\displaystyle K_{x}\in \mathbb {R} ^{n\times n}} and K y ∈ R n × n {\displaystyle K_{y}\in \mathbb {R} ^{n\times n}} , KCCA seeks coefficients α {\displaystyle \alpha } and β {\displaystyle \beta } that maximize:

ρ = max α , β α ■ K x K y β α ■ K x 2 α β ■ K y 2 β {\displaystyle \rho =\max _{\alpha ,\beta }{\frac {\alpha ^{\top }K_{x}Ky\beta }{{\sqrt {\alpha ^{\top }K_{x}^{2}\alpha }}{\sqrt {\beta ^{\top }K_{y}^{2}\beta }}}}}

To prevent overfitting , regularization terms are typically added, resulting in:

ρ = max α , β α T K x K y β α T (K x 2 + λ x K x) α β T (K y 2 + λ y K y) β {\displaystyle \rho =\max _{\alpha ,\beta }{\frac {\alpha ^{T}K_{x}K_{y}\beta }{{\sqrt {\alpha ^{T}\left(K_{x}^{2}+\lambda _{x}K_{x}\right)\alpha }}{\sqrt {\;\beta ^{T}\left(K_{y}^{2}+\lambda _{y}K_{y}\right)\beta }}}}}

where λ x {\displaystyle \lambda _{x}} and λ y {\displaystyle \lambda _{y}} are regularization parameters. KCCA has proven effective for tasks such as cross-modal retrieval and semantic analysis, though it faces computational challenges with large datasets due to its O (n 2) {\displaystyle O(n^{2})} memory requirement for sorting kernel matrices.

KCCA was proposed independently by several researchers. [5] [6] [7] [8]

Deep CCA

Deep canonical correlation analysis (DCCA), introduced in 2013, employs neural networks to learn nonlinear transformations for maximizing the correlation between modalities. [1] DCCA uses separate neural networks f x {\displaystyle f_{x}} and f y {\displaystyle f_{y}} for each modality to transform the original data before applying CCA:

max W x , W y , θ x , θ y corr ■ (f x (X ; θ x) , f y (Y ; θ y)) {\displaystyle \max _{W_{x},W_{y},\theta _{x},\theta _{y}}\operatorname {corr} \left(f_{x}(X;\theta _{x}),f_{y}(Y;\theta _{y})\right)}

where θ x {\displaystyle \theta _{x}} and θ y {\displaystyle \theta _{y}} represent the parameters of the neural networks, and W x {\displaystyle W_{x}} and W y {\displaystyle W_{y}} are the CCA projection matrices. The correlation objective is computed as:

corr ■ (H x , H y) = tr ■ (T − 1 / 2 H x T H y S − 1 / 2) {\displaystyle \operatorname {corr} (H_{x},H_{y})=\operatorname {tr} \left(T^{-1/2}H_{x}^{T}H_{y}S^{-1/2}\right)}

where H x = f x (X) {\displaystyle H_{x}=f_{x}(X)} and H y = f y (Y) {\displaystyle H_{y}=f_{y}(Y)} are the network outputs, T = H x T H x + r x I {\displaystyle T=H_{x}^{T}H_{x}+r_{x}I} , S = H y T H y + r y I {\displaystyle S=H_{y}^{T}H_{y}+r_{y}I} and r x , r y {\displaystyle r_{x},r_{y}} are the regularization parameters. DCCA overcomes the limitations of linear CCA and kernel CCA by learning complex nonlinear relationships while maintaining computational efficiency for large datasets through mini-batch optimization. [9]

Graph-based methods

Graph-based approaches for multimodal representation learning leverage graph structure to model relationships between entities across different modalities. These methods typically represent each modality as a graph and then learn embedding that preserve cross-modal similarities, enabling more effective joint representation of heterogeneous data. [10]

One such method is cross-modal graph neural networks (CMGNNs) that extend traditional graph neural networks (GNNs) to handle data from multiple modalities by constructing graphs that capture both intra-modal and inter-modal relationships. These networks model interactions across modalities by representing them as nodes and their relationships as edges. [11]

Other graph-based methods include Probabilistic Graphical Models (PGMs) such as deep belief networks (DBN) and deep Boltzmann machines (DBM). These models can learn a joint representation across modalities, for instance, a multimodal DBN achieves this by adding a shared restricted Boltzmann Machine (RBM) hidden layer on top of modality-specific DBNs. [1] Additionally, the structure of data in some domains like Human-Computer Interaction (HCI), such as the view hierarchy of app screens, can potentially be modeled using graph-like structures. The field of graph representation learning is also relevant, with ongoing progress in developing evaluation benchmarks. [12]

Diffusion maps

Another set of methods relevant to multimodal representation learning are based on diffusion maps and their extensions to handle multiple modalities.

Multi-view diffusion maps

Multi-view diffusion maps address the challenge of achieving multi-view dimensionality reduction by effectively utilizing the availability of multiple views to extract a coherent low-dimensional representation of the data. The core idea is to exploit both the intrinsic relations within each view and the mutual relations between the different views, defining a cross-view model where a random walk process implicitly hops between objects in different views. A multi-view kernel matrix is constructed by combining these relations, defining a cross-view diffusion process and associated diffusion distances. The spectral decomposition of this kernel enables the discovery of an embedding that better leverages the information from all views. This method has demonstrated utility in various machine learning tasks, including classification, clustering, and manifold learning. [13]

Alternating diffusion

Alternating diffusion based methods provide another strategy for multimodal representation learning by focusing on extracting the common underlying sources of variability present across multiple views or sensors. These methods aim to filter out sensor-specific or nuisance components, assuming that the phenomenon of interest is captured by two or more sensors. The core idea involves constructing an alternating diffusion operator by sequentially applying diffusion processes derived from each modality, typically through their product or intersection. This process allows the method to capture the structure related to common hidden variables that drive the observed multimodal data. [14]

See also

Representation learning

Canonical correlation

Deep learning

Multimodal learning

Nonlinear dimensionality reduction

References