Title: Multimodal learning

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

PageID: 46975535

Categories: Category:Artificial neural networks, Category:Multimodal interaction

-----

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

Multimodal learning is a type of deep learning that integrates and processes multiple types of data, referred to as modalities , such as text, audio, images, or video. This integration allows for a more holistic understanding of complex data, improving model performance in tasks like visual question answering, cross-modal retrieval, [1] text-to-image generation, [2] aesthetic ranking, [3] and image captioning. [4]

Large multimodal models, such as Google Gemini and GPT-4o , have become increasingly popular since 2023, enabling increased versatility and a broader understanding of real-world phenomena. [5]

Motivation

Data usually comes with different modalities which carry different information. For example, it is very common to caption an image to convey the information not presented in the image itself. Similarly, sometimes it is more straightforward to use an image to describe information which may not be obvious from text. As a result, if different words appear in similar images, then these words likely describe the same thing. Conversely, if a word is used to describe seemingly dissimilar images, then these images may represent the same object. Thus, in cases dealing with multi-modal data, it is important to use a model which is able to jointly represent the information such that the model can capture the combined information from different modalities.

Multimodal transformers

Transformers can also be used/adapted for modalities (input or output) beyond just text, usually by finding a way to "tokenize" the modality.

Multimodal models can either be trained from scratch, or by finetuning. A 2022 study found that Transformers pretrained only on natural language can be finetuned on only 0.03% of parameters and become competitive with LSTMs on a variety of logical and visual tasks, demonstrating transfer learning . [6] The LLaVA was a vision-language model composed of a language model (Vicuna-13B) [7] and a vision model (ViT -L/14), connected by a linear layer. Only the linear layer is finetuned. [8]

Vision transformers [9] adapt the transformer to computer vision by breaking down input images as a series of patches, turning them into vectors, and treating them like tokens in a standard transformer.

Conformer [10] and later Whisper [11] follow the same pattern for speech recognition , first turning the speech signal into a spectrogram , which is then treated like an image, i.e. broken down into a series of patches, turned into vectors and treated like tokens in a standard transformer.

Perceivers [12] [13] are a variant of Transformers designed for multimodality.

Multimodal large language models

Multimodality means having multiple modalities, where a " modality " refers to a type of input or output, such as video, image, audio, text, proprioception , etc. [19] For example, Google PaLM model was fine-tuned into a multimodal model and applied to robotic control . [20] LLaMA models have also been turned multimodal using the tokenization method, to allow image inputs, [21] and video inputs. [22] GPT-4o can process and generate text, audio and images. [23] Such models are sometimes called large multimodal models (LMMs). [24]

A common method to create multimodal models out of an LLM is to "tokenize" the output of a trained encoder. Concretely, one can construct an LLM that can understand images as follows: take a trained LLM, and take a trained image encoder E {\displaystyle E} . Make a small multilayered perceptron f {\displaystyle f} , so that for any image y {\displaystyle y} , the post-processed vector f (E (y)) {\displaystyle f(E(y))} has the same dimensions as an encoded token. That is an "image token". Then, one can interleave text tokens and image tokens. The compound model is then fine-tuned on an image-text dataset. This basic construction can be applied with more sophistication to improve the model. The image encoder may be frozen to improve stability. [25] This type of method, where embeddings from multiple modalities are fused and the predictor is trained on the combined embeddings, is called early fusion.

Multimodal deep Boltzmann machines

A Boltzmann machine is a type of stochastic neural network invented by Geoffrey Hinton and Terry Sejnowski in 1985. Boltzmann machines can be seen as the stochastic , generative counterpart of Hopfield nets . They are named after the Boltzmann distribution in statistical mechanics. The units in Boltzmann machines are divided into two groups: visible units and hidden units. Each unit is like a neuron with a binary output that represents whether it is activated or not. [28] General Boltzmann machines allow connection between any units. However, learning is impractical using general Boltzmann Machines because the computational time is exponential to the size of the machine [citation needed] . A more efficient architecture is called restricted Boltzmann machine where connection is only allowed between hidden unit and visible unit, which is described in the next section.

Multimodal deep Boltzmann machines can process and learn from different types of information, such as images and text, simultaneously. This can notably be done by having a separate deep Boltzmann machine for each modality, for example one for images and one for text, joined at an additional top hidden layer. [29]

Applications

Multimodal machine learning has numerous applications across various domains:

Cross-modal retrieval : cross-modal retrieval allows users to search for data across different modalities (e.g., retrieving images based on text descriptions), improving multimedia search engines and content recommendation systems. Models like CLIP facilitate efficient, accurate retrieval by embedding data in a shared space, demonstrating strong performance even in zero-shot settings. [30]

Classification and missing data retrieval : multimodal Deep Boltzmann Machines outperform traditional models like support vector machines and latent Dirichlet allocation in classification tasks and can predict missing data in multimodal datasets, such as images and text.

Healthcare diagnostics : multimodal models integrate medical imaging, genomic data, and patient records to improve diagnostic accuracy and early disease detection, especially in cancer screening. [31] [32] [33]

Content generation : models like DALL·E generate images from textual descriptions, benefiting creative industries, while cross-modal retrieval enables dynamic multimedia searches. [34]

Robotics and human-computer interaction : multimodal learning improves interaction in robotics and AI by integrating sensory inputs like speech, vision, and touch, aiding autonomous systems and human-computer interaction .

Emotion recognition : combining visual, audio, and text data, multimodal systems enhance sentiment analysis and emotion recognition , applied in customer service, social media, and marketing.

See also

Hopfield network

Markov random field

Markov chain Monte Carlo

References