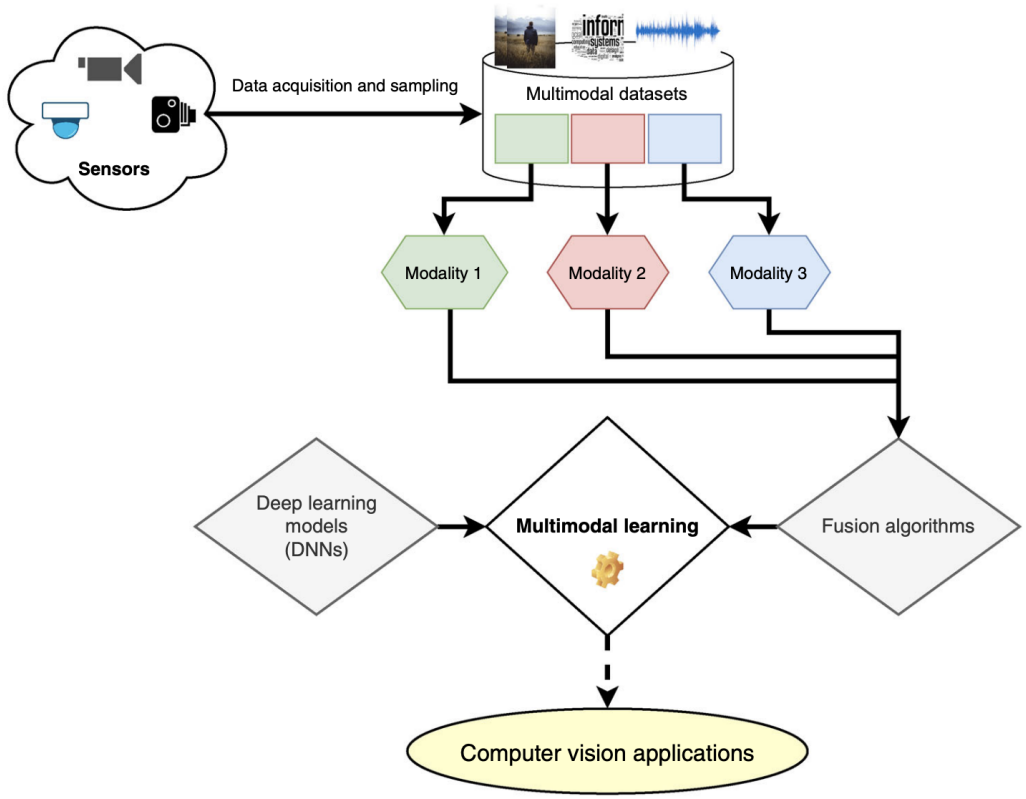
A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets

Authors: Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, Abdellatif Mtibaa
May 15, 2021

Introduction

Multimodal learning is an important new area of research which aims to perform deep learning on a task which involves multiple modalities. Different modes include text, images, audio, and other signals which are input datasets. In particular, this paper focuses on multimodal data involving images and related vision tasks. The paper focuses on downstream tasks such as multimodal data representation, multimodal fusion, multitask learning, multimodal alignment, multimodal transfer learning, and zero-shot learning. It also surveys a list of benchmark datasets and provides an overview of algorithms and techniques used for multimodal deep learning on vision tasks.

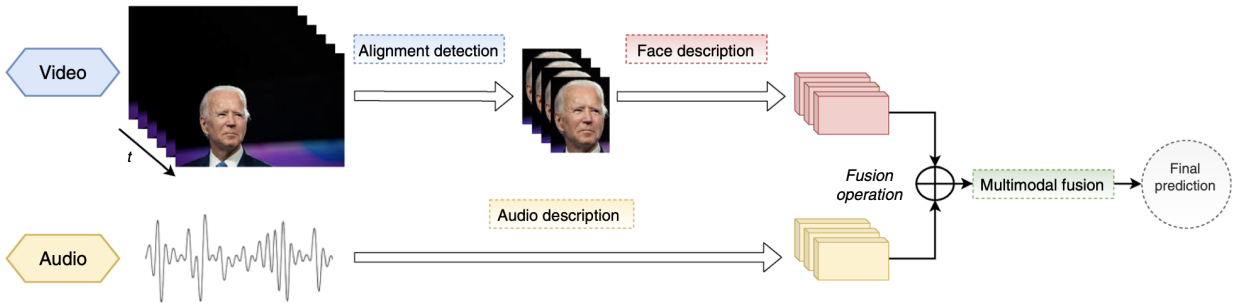


Challenges

There are many challenges with using multimodal data. Firstly, multimodal data is often high dimensional which makes it more difficult to learn cleaner representations and wrangle the data. Labelled multimodal training data is often not available. It can also be difficult to scale real time systems to process multimodal data.

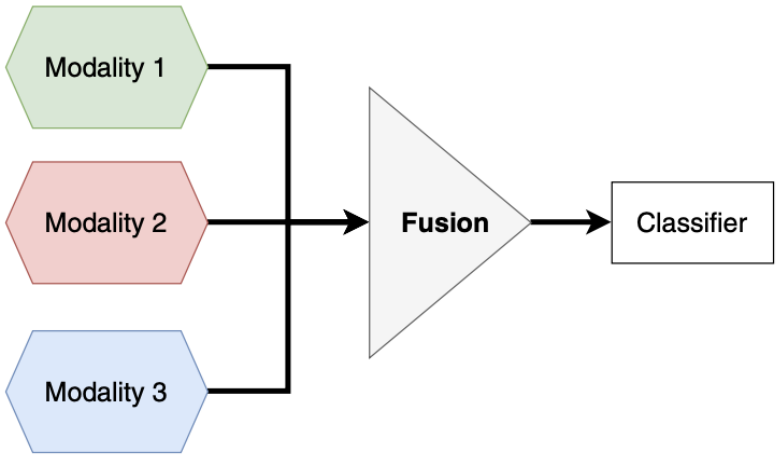
Multimodal Fusion Algorithms

Up until recently, most applications using deep learning models have been using monomodal models which learn from a single type of data stream. Models would learn the vector representations of a single data of a single type. On the other hand, a multimodal model would aggregate the representations of multiple data modalities during the learning stage. Until recently, applications would train separate monomodal models and join the results after obtaining the representations or results of the monomodal model through a multimodal fusion.

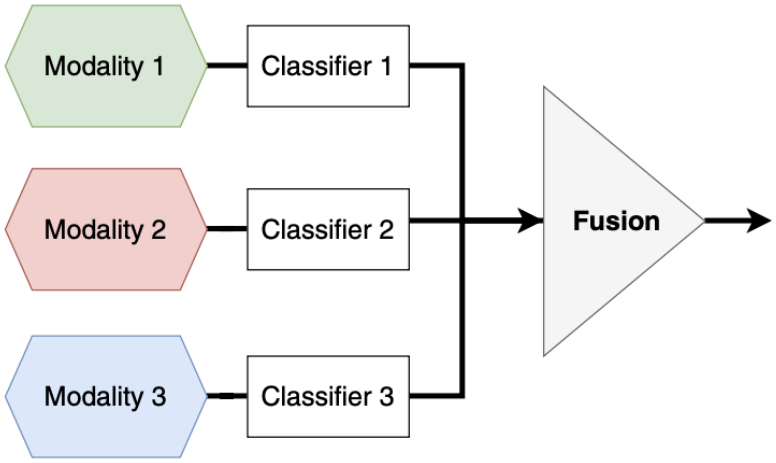


There are different approaches to fusion:

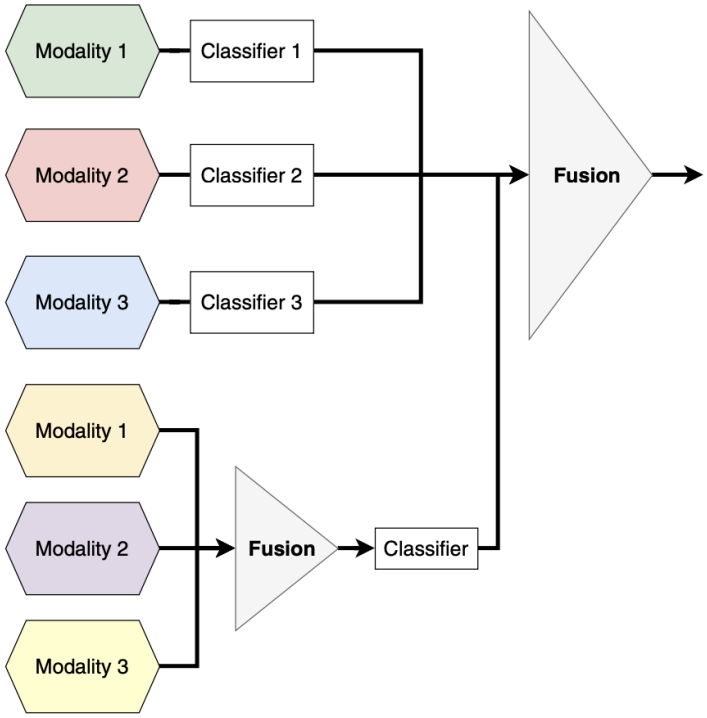
1. Early fusion: low level features from each modality are fused before classification



2. Late fusion: classify features from each modality before fusing



3. Hybrid fusion: combine multimodal features from early and late fusion before classification



Past Conventional Approaches

The past conventional approaches include

1. Kernel based
2. Graphical model based (ie. Hidden Markov Model)
3. Correlation analysis based
4. Deep learning based (including autoencoders, CNNs, RNNs, GANs, attention-based)

Modern Approaches

Multitask Learning

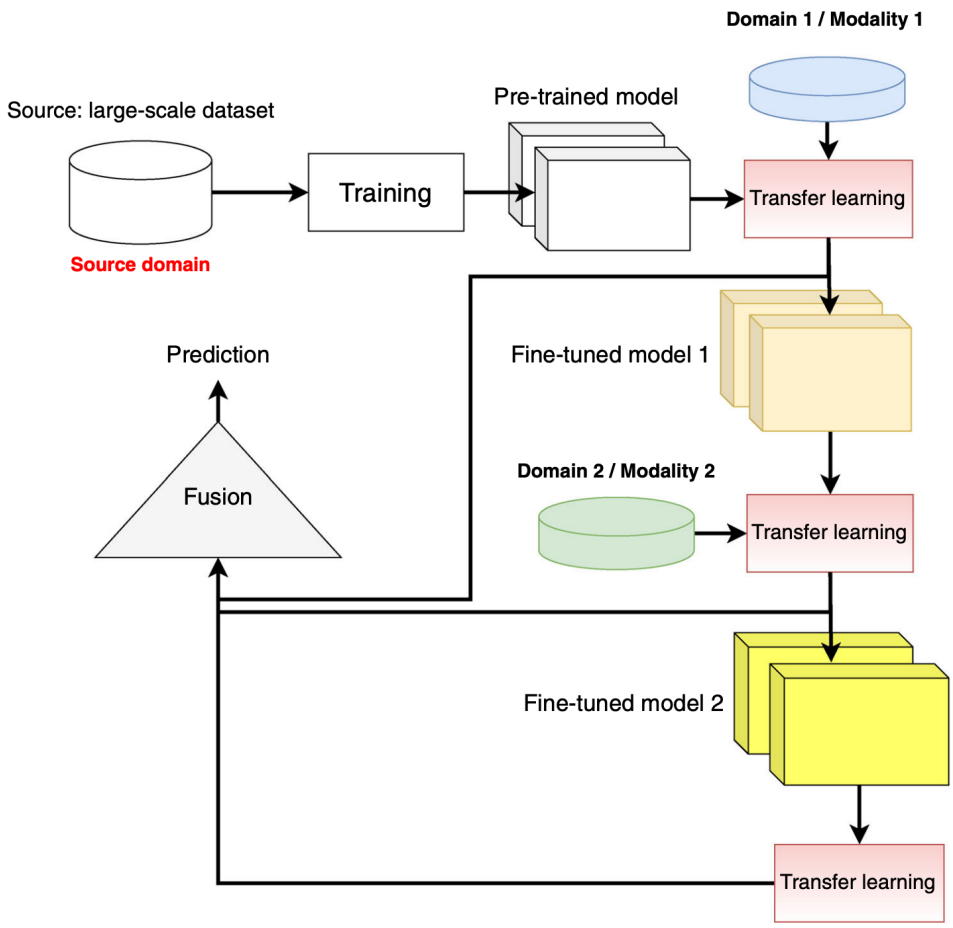
With multitask learning (MTL), the model learns a shared representation that can be used for several tasks and allows for better model generalizability. MTL can perform either soft or hard sharing of parameters. With soft sharing, the model extracts features and simultaneously learns similarity relationships between them. With hard sharing, a more generic feature representation is extracted for different tasks using the same parameters.

Multimodal Alignment

Multimodal alignment is when the features of multiple different modalities are linked together through structural or spatial information. In the case of images, a spatial mapping must be aligned for the image and other sensor data.

Multimodal Transfer Learning

Transfer learning is when we leverage a large pre-trained model that has been trained already on millions of samples of data, and fine-tune from that model as a baseline. Fine-tuning speeds up the overall training process and generally leads to improved accuracy and quality. If there is a multimodally trained model on particular modalities of interests, we could leverage that for a related but different task using the same modalities.



Zero-shot Learning

Zero-shot learning is when there is a lack of labelled training examples for a model to learn all possible labels well. This approach tends to also attempt to generate synthetic samples of previously unseen classes through usage of GANs.

Vision Tasks and Applications

Object Detection

Objects can be detected using a variety of modalities including vision, external sensors, thermal data, and audio. More recently new signals have been incorporated into object detection applications such as depth perception, optical flow, and LiDAR point clouds. Shared representations can be learned to best improve the classification performance of models.

Human and face recognition in particular becomes a multimodal problem when additional data such as biometric data and face reconstruction data.

Semantic Segmentation

Image segmentation can be aided with additional modalities of data such as soft correspondances, 3D scenes, and temperature information.

Image Retrieval

Query by image content is a popular search engine task. Text and image multimodal deep learning representations are a popular application for this task.

Image Captioning

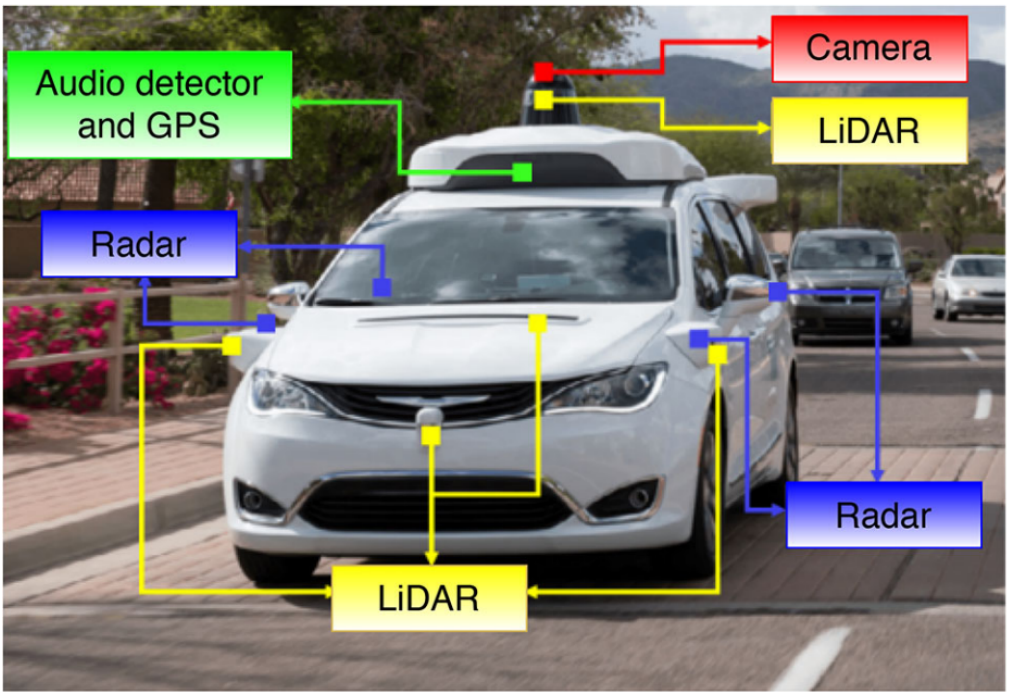
Images can have a textual caption generated for the image. In the past, multimodal models encoded both the image and the text using CNNs and RNNs to learn the representations. Then a multimodal model would decode these two parts into a caption similar to the image. As for video captioning, the data used could include temporal data, audio data, and motion data.

Medical Data Analysis

Recently, more modalities of medical data are taken to perform classification or other anomaly detection tasks for medical purposes. Data such as multiple different images can be used such as x-rays and CT scans allow for early detection of conditions.

Autonomous systems

Autonomous vehicles in particular utilize image, depth, and LiDAR systems to combine different modalities in order to operate. Mobile robots also utilize other sensors to detect their environment, making multimodal systems necessary.



Vision Multimodal Datasets

Several popular multimodal vision datasets were created to help benchmark new multimodal model performance.

- RGB-D Object: 300 objects from 51 categories and multiple view angles, 250,000 samples
- BigBIRD: 125 objects, 600 RGB-D point clouds, 600 images taken by a Kinect and DSLR camera
- RGB-D Semantic Segmentation: 3D models of objects in six categories such as juice, bottles, coffee cans and salt
- RGB-D Scenes: video scenes
- RGB-D People: 3000 images from Kinect camera
- SceneNet RGB-D: 5 million RGB-D images
- Kinetics-400: 400 classes of human actions from 300,000 Youtube videos
- MPI-Sintel: 1,040 long sequences with vision and optical flow data

Conclusion

Multimodal data is challenging to use although it has many relevant use-cases for important applications including medical, autonomous systems, and object detection. There has been a lot of recent research to advance multimodal learning including multitask learning, transfer learning, and zero-shot learning in the computer vision space. The other modalities of data used often include text, sensor data, depth perception data, and spatio-temporal signals. There are a

variety of tasks which these modalities and techniques apply to, and more research will surely be done in these domains.

References

All images were taken from the paper (https://doi.org/10.1007/s00371-021-02166-7).