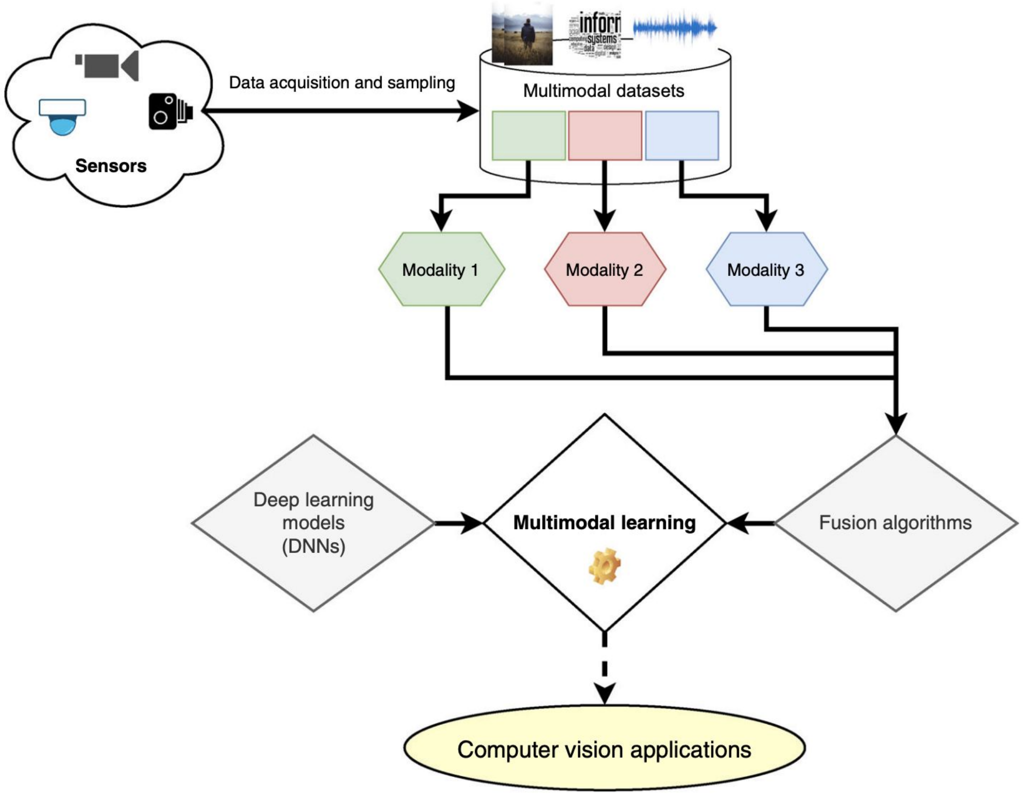
A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets

Shannon Phu

Based on Bayoudh, Knani, Hamdaoui, Mtibaa

Introduction

- Important new area of research which aims to perform deep learning on a task which involves multiple modalities
- Different modes include text, images, audio, sensors, and other signals
- This paper focuses on multimodal data involving images and related vision tasks

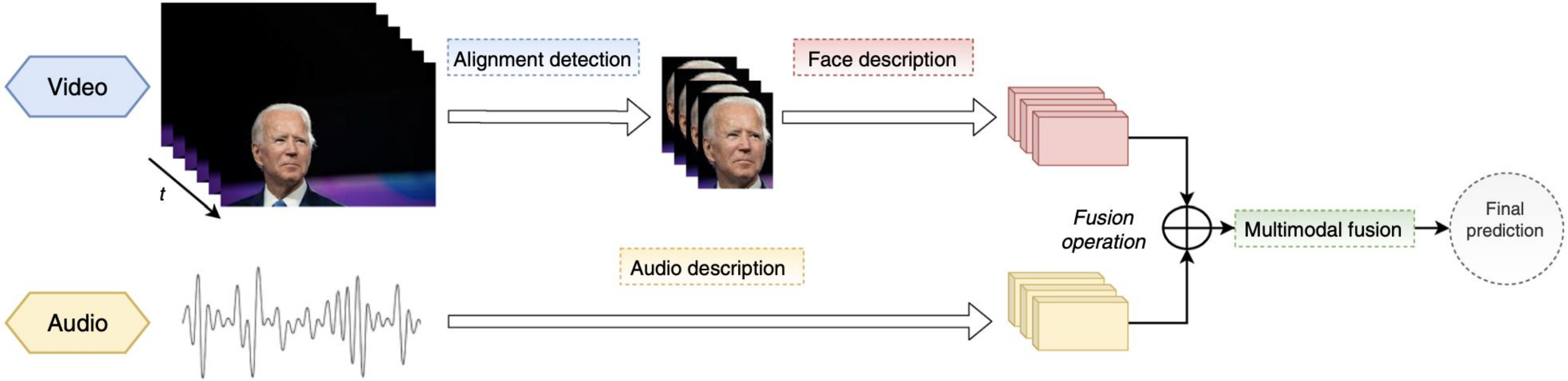


Challenges

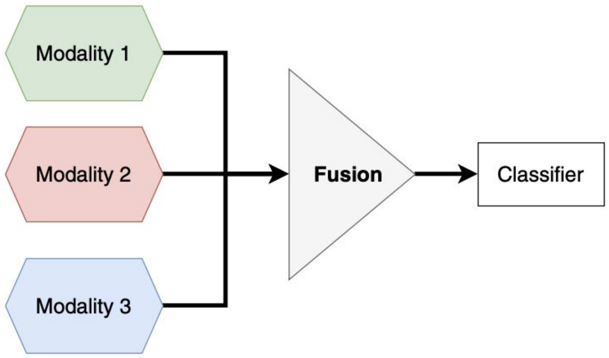
- High dimensional which makes it more difficult to learn cleaner representations and wrangle the data
- Labelled multimodal training data is often not available
- Can also be difficult to scale real time systems to process multimodal data

Multimodal Fusion Algorithms

- multimodal model would aggregate the representations of multiple data modalities during the learning stage

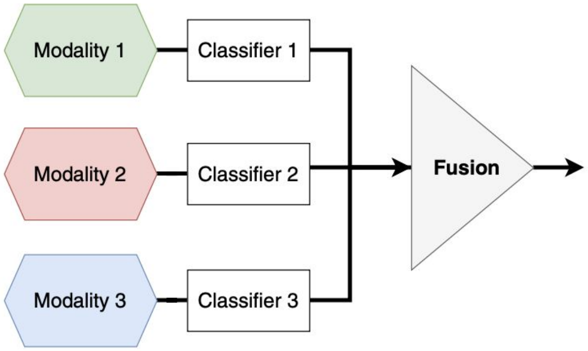


Early fusion



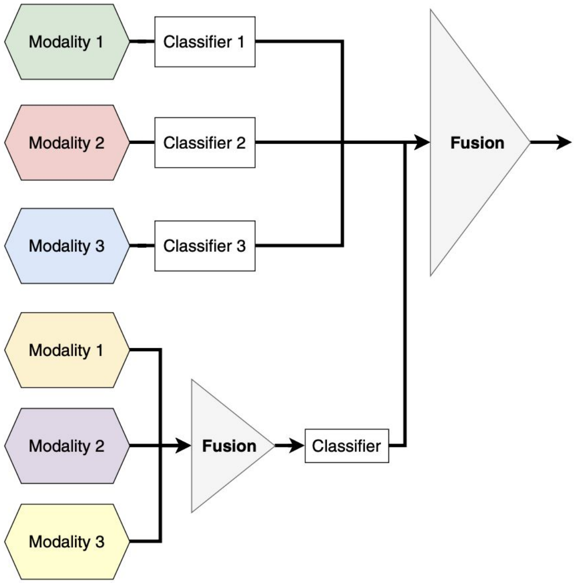
low level features from each modality are fused before classification

Late fusion



classify features from each modality before fusing

Hybrid fusion



combine multimodal features from early and late fusion before classification

Past Conventional Approaches

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
  - 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
  - features of multiple different modalities are linked together through structural or spatial information
- Multimodal Transfer Learning
  - leverage a large pre-trained model trained on multimodal data
- Zero-shot Learning
  - when there is a lack of labelled training examples for a model to learn all possible labels well
  - generate synthetic samples of previously unseen classes through usage of GANs

Vision Tasks and Applications

- Object Detection
  - variety of modalities including vision, external sensors, thermal data, audio, depth perception, optical flow, and LiDAR point clouds
  - Human and face recognition in particular becomes a multimodal problem when additional data such as biometric data and face reconstruction data
- Semantic Segmentation
  - additional modalities of data such as soft correspondances, 3D scenes, and temperature information
- Image Retrieval
  - Text and image multimodal data
- Image Captioning
  - multimodal models encoded both the image and the text using CNNs and RNNs to learn the representations
  - video captioning, the data used could include temporal data, audio data, and motion data
- Medical Data Analysis
  - multiple different images can be used such as x-rays and CT scans allow for early detection of conditions
- Autonomous systems
  - image, depth, and LiDAR systems to combine different modalities for AVs
  - Mobile robots also utilize other sensors to detect their environment

Vision Multimodal Datasets

- 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.