Published in Haileleol Tibebu

Haylat T Follow

Jan 30, 2020 · 7 min read · Listen

Save

INTRODUCTION TO DATA FUSION

multi-modality

One of the most famous machine learning algorithms, neural networks, have been playing a significant role in recent years due to their ability to train with high accuracy. A neural network is a human brain inspired deep learning method. Deep learning has become a prominent research interest in both academia and industries, mainly because of its high performance compared to other machine learning architecture.

Deep learning in a single domain dataset has been successful. Current research involves multimodal input data. Lahal et al. [3]defines multimodality as a system which observed by multiple sensors. The aim of using multimodality is to extract and mix important information from individual sensors and use this mixed feature to solve a given problem. Thus, the expected output will have a richer representation and performance than the individual modalities. Multimodal data analysis is a practical solution to several field of studies like, Medicine, business and driverless technology and gaming. Common remote sensing apparatus like camera, LIDAR, radar and ultrasonic are often fused [4].

Multimodality techniques

There are three techniques used for multimodal data fusion[5] [6].

1. Early fusion or data-level fusion

Data level fusion is a traditional way of fusing multiple data before conducting the analysis (Figure 3). This method is referred to as input level fusion. Research [6] proposes two possible approaches for early fusion technique. The first approach is combining data by removing the correlation between two sensors. The second approach is to fuse data at its lower dimensional common space. There are many statistical solutions which can be used to accomplish one or both methods, including principal component analysis (PCA), canonical correlation analysis and independent component analysis.

Early fusion is applicable on raw data or pre-processed data obtained from sensors. Data features should be extracted from the data before fusion, otherwise the process will be challenging especially when the data sources have different sampling rates between the modalities. Synchronization of data sources is also challenging when one data source is discrete and the others are continuous. Hence, converting data sources into a single feature vector is a significant challenge in early data fusion.

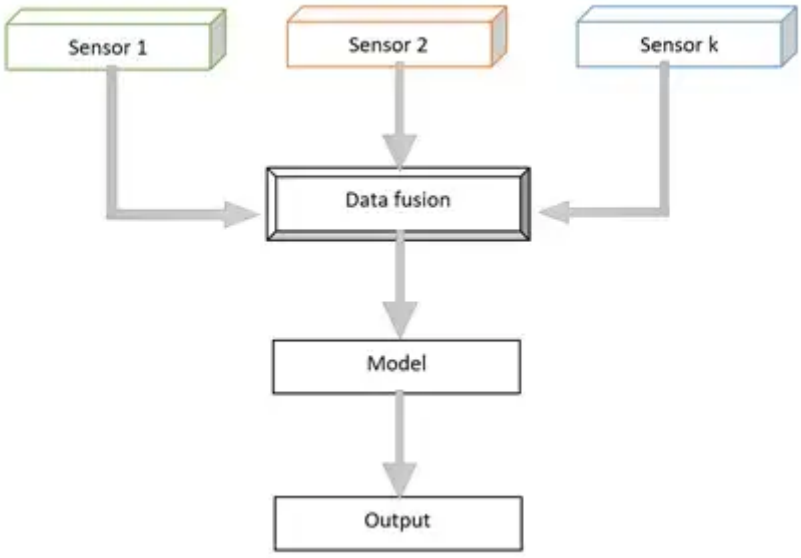

Fig 1. Early fusion or data-level fusion

The assumption behind early data fusion is the conditional independence between multiple data sources. According to Sebe et al [7], this assumption is not always true as multiple modalities can have highly correlated features, for example video and depth cues. Another paper [8] also states that different modalities can contain information that are correlated to each other at a higher level. Thus, the output of each modality can be assumed to be processed independently of one another. Poria et al [9] implemented early stage data fusion which involved concatenation of the

features in a multimodal stream, this can be assumed as the simplest form of early stage data fusion.

There are two disadvantages in using early stage data fusion. One of the main disadvantages of this method is that a large amount of data will be deducted from the modalities to make a common ground before fusion. Once the data have common matrices, they are analysed using a machine learning algorithm. The other disadvantage of this method is synchronizing the timestamp of the different modalities. A common way to overcome this disadvantage is to collect the data or signals at a common sampling rate. Other mitigating solutions are proposed by Martinez et al[10] which includes training, pooling and convolution fusion. These proposed methods were achieved by the fusion of sequential discrete events with continuous data.

2. Late fusion or decision level fusion

Late fusion uses data sources independently followed by fusion at a decision-making stage (Figure 4). Late data fusion is inspired by the popularity of ensemble classifiers [11]. This technique is much simpler than the early fusion method, particularly when the data sources are significantly varied from each other in terms of sampling rate, data dimensionality and unit of measurement. Late fusion often gives better performance because errors from multiple models are dealt with independently — thus error are uncorrelated. However, Ramachandram et al [12] argue that there is no conclusive evidence that late fusion performs better than early fusion. Yet, a number of researchers use late or decision level fusion to analyse multimodal data problems [13][14][15].

Different rules exist to determine the optimal way of deciding how to finally combine each of the independently trained models. Bayes rules, max-fusion and average-fusion are some of the commonly late fusion rules.

When the input data streams are significantly varied in terms of dimensionality and sampling rate, using late fusion is a simpler and more flexible approach.

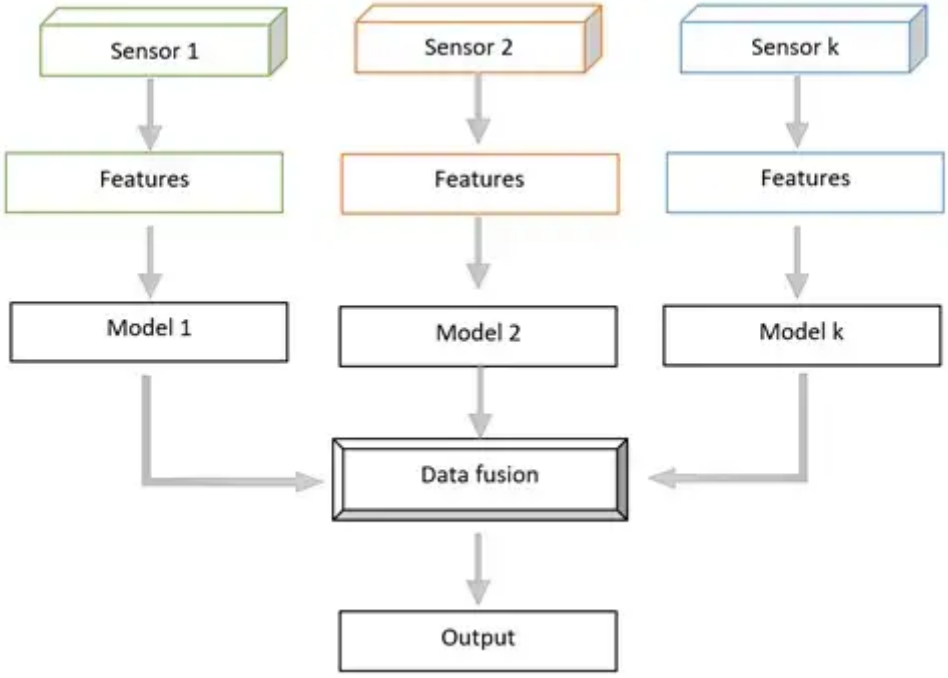

Figure 2. Late fusion or decision fusion

3. Intermediate fusion

The architecture of intermediate fusion is built on the basis of the popular deep neural network. This method is the most flexible method, allowing for data fusion at different stages of model training. Neural network based multimodal data fusion has greatly improved performance.

Intermediate fusion changes input data into a higher level of representation (features) through multiple layers. Each individual layer operates linear and nonlinear functions which transform the input data's scale, skew and swing and gives a new representation of the original input data. Intermediate fusion in a deep learning multimodal context is a fusion of different modalities representations into a single hidden layer so that the model learns a joint representation of each of the modalities. Features can be learned from different kinds of layers including: 2D convolution, 3D convolution and fully connected. The layer where the fusion of different modality features has taken place is called a fusion layer or a shared representation layer.

Different modalities can be fused simultaneously into a single shared representation layer or this can be performed gradually using one or multiple modalities at a time (Figure 5). Although it is possible to fuse multiple modality features or weights in a single layer, it may lead to model overfitting or the network may fail to learn the relationship between each modality.

One method to improve deep multimodal fusion performance is to reduce the dimensionality of the data. Li et al [16] use principal component analysis (PCA) and Ding et al [17] use autoencoders to reduce dimensionality of the network after constructing a fusion layer or shared representation layer. As opposed to early level fusion and late fusion, intermediate fusion offers flexibility to fuse features at different depths.

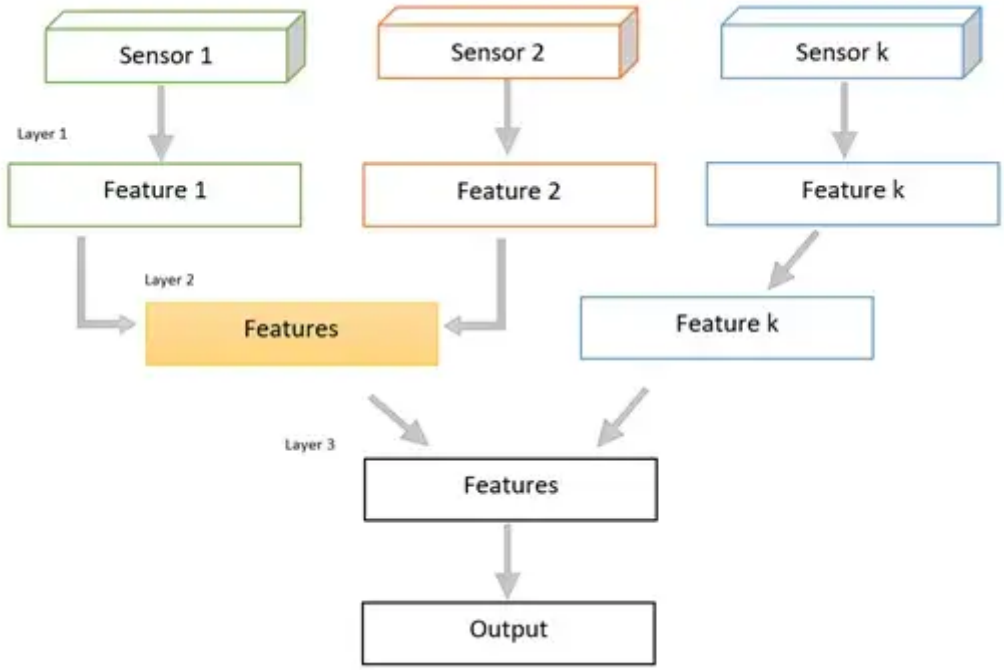

Figure 3. intermediate fusion

A research paper by Karpathy et al [18] uses a “slow-fusion” network where training video stream features are gradually fused across multiple fusion layers. This approach performs better in a large-scale video stream classification problem. Other similar research [19] shows a gradual fusion method which fused highly correlated input modalities first and less correlated ones progressively after (i.e. visual input modalities, then motion input modal then audio input modalities). This paper proposed a state-of-the-art performance in communicative gesture recognition.

Reference

[3] *D. Lahat, T. Adali, and C. Jutten, “Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects,” Proceedings of the IEEE. 2015.*

[4] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[5] D. Lahat et al., "Multimodal Data Fusion : An Overview of Methods , Challenges and Prospects To cite this version : HAL Id : hal-01179853 Multimodal Data Fusion : An Overview of Methods , Challenges and Prospects," arXiv, vol. 103, no. 9, pp. 1–26, 2015.

[6] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," Inf. Fusion, 2013.

[7] Machine Learning in Computer Vision. 2005.

[8] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.

[9] S. Poria, E. Cambria, and A. Gelbukh, "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis," no. September, pp. 2539–2544, 2015.

[10] H. P. Martínez and G. N. Yannakakis, "Deep Multimodal Fusion," 2014.

[11] L. I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms," Wiley, 2004.

[12] R. Dhanesh and T. Graham W, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," IEEE Signal Process. Mag., vol. 34, no. 6, pp. 96–108, 2017.

[13] Z. Simonyan, Karen and Andrew, "Two-Stream convolutional networks for Action Recognition," in NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems — Volume 1, 2004.

[14] D. Wu et al., "Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition," IEEE Trans. Pattern Anal. Mach. Intell., 2016.

[15] S. E. Kanou et al., "Combining modality specific deep neural networks for emotion recognition in video," 2013.

[17] C. Ding and D. Tao, "Robust Face Recognition via Multimodal Deep Face Representation," IEEE Trans. Multimed., 2015.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014.

[19] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," IEEE Trans. Pattern Anal. Mach. Intell., 2016.