Artificial Intelligence Review (2023) 56:6661–6704
https://doi.org/10.1007/s10462-022-10332-z

Deep learning on multi-view sequential data: a survey

Zhuyang Xie[1,2] · Yan Yang[1,2] · Yiling Zhang[1,2] · Jie Wang[1,2] · Shengdong Du[1,2]

Published online: 29 November 2022

Abstract
With the progress of human daily interaction activities and the development of industrial society, a large amount of media data and sensor data become accessible. Humans collect these multi-source data in chronological order, called multi-view sequential data (MvSD). MvSD has numerous potential application domains, including intelligent transportation, climate science, health care, public safety and multimedia, etc. However, as the volume and scale of MvSD increases, the traditional machine learning methods become difficult to withstand such large-scale data, and it is no longer appropriate to use hand-craft features to represent these complex data. In addition, there is no general framework in the process of mining multi-view relationships and integrating multi-view information. In this paper, We first introduce four common data types that constitute MvSD, including point data, sequence data, graph data, and raster data. Then, we summarize the technical challenges of MvSD. Subsequently, we review the recent progress in deep learning technology applied to MvSD. Meanwhile, we discuss how the network represents and learns features of MvSD. Finally, we summarize the applications of MvSD in different domains and give potential research directions.

Keywords Deep neural networks · Multi-view · Sequential data · Spatio-temporal

✉ Yan Yang
yyang@swjtu.edu.cn

Zhuyang Xie
zyxie@my.swjtu.edu.cn

Yiling Zhang
zylscience@foxmail.com

Jie Wang
JackWang@my.swjtu.edu.cn

Shengdong Du
sddu@swjtu.edu.cn

[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China

[2] Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory, Southwest Jiaotong University, Chengdu 611756, China

1 Introduction

With the development of social media, more and more human activities have become public and accessible. In addition, the application of a large number of acquisition equipments and sensors has made it easier for us to obtain information about the surrounding world. Humans collect these multi-source data in chronological order and obtain multi-view sequential data (MvSD). MvSD has broad researches and applications in various domains, including smart transportation, climate science, social media, health care, crime analysis, etc. However, as the volume and scale of MvSD increases, classical data mining methods are no longer applicable. On the one hand, the construction of hand-craft features is restricted by limited human knowledge, thus conventional methods are difficult to represent such complex data. On the other hand, MvSD changes dynamically over time and presents self-correlated, the traditional machine learning methods can not fully mine the knowledge mechanism in sequential data and it is difficult to effectively analyze the hidden attributes. Simultaneously, MvSD collected from various domains or obtained from diverse sensors leads to heterogeneity among views. Thus, how to make full use of the diversity among different views and fuse the latent knowledge in MvSD has attracted extensive research.

In recent years, deep learning has swept many fields and achieved remarkable achievements, such as object detection (Girshick 2015; Ren et al. 2015; He et al. 2016; Redmon et al. 2016; Liu et al. 2016a), image segmentation (Long et al. 2015; Ronneberger et al. 2015; Lin et al. 2017; He et al. 2017; Zhao et al. 2017c; Badrinarayanan et al. 2017), natural language processing (Kiros et al. 2014; Bahdanau et al. 2014; Cheng et al. 2016; Vaswani et al. 2017), etc. Deep learning has brought the possibility to solve the above problems with its general data understanding and parallel computing capabilities. First of all, the superiority of deep learning is based on its feature extraction ability, which breaks the performance of human-engineered features through end-to-end learning. Among them, convolutional neural network (CNN) achieves excellent performance on regular raster data, while recurrent neural network (RNN) is adapted to sequence data and model the correlations. Second, classical methods based on small datasets are untenable in MvSD. In contrast, the performance of deep models will be further improved with massive data samples. Therefore, deep models perform better feature representation and learning on larger datasets. Third, conventional machine learning methods generally exploits some linear functions to fit latent data structure and is not able to express complex models. As we all know, deep networks have nonlinear approximation capabilities and learn rules by optimizing the loss function as much as possible.

It is worth mentioning that MvSD is collected from multiple domains. For these multi-source sequential data, the analysis method based on multi-view learning has more sufficient feature representation capability than single-view learning. Multi-view deep learning can not only be used to analyze the implicit feature correlation and internal dynamic changes in sequential data, but also help solve the incompleteness and uncertainty in sequential data analysis. Taking video sentiment analysis as an example. It usually consists of three kinds of data: text sentences, images and audio clips. It is not comprehensive to analyze the expression only in text streams, because language is ambiguous in different situations. By combining facial features and speakers pronunciation, the speakers attitude is more accurately inferred. As for the traffic flow forecasting task, some external factors will also affect the prediction performance, such as weather, holidays and events. Injecting these factors into the network will further assist in prediction.

In the past few decades, a large number of machine learning techniques have been used for multi-view data, resulting in multi-view representation, multi-view clustering, multi-view fusion, etc. With the large amount of multimedia data available in recent years, multi-view learning has become a promising research. Benefiting from a large number of researches and sufficient theories of multi-view learning (Khan et al. 2022a; Yin and Sun 2019), some deep learning-based multi-view algorithms are gradually being studied (Yin and Sun 2019; Sun and Zong 2020; Mao and Sun 2020), forming deep multi-view learning (Yao et al. 2018; Wang et al. 2015; Kan et al. 2016; Sun et al. 2020d) and deep multi-view clustering (Li et al. 2019; Khan et al. 2022b; Xia et al. 2022). For example, Deep CCA (Andrew et al. 2013), which is extended from canonical correlation analysis (CCA) (Hotelling 1992), learns non-linear mappings between different views through stacked multi-layer neural networks. Deep matrix factorization (Zhao et al. 2017b; Huang et al. 2020a), which applies non-negative matrix factorization (NMF) from traditional clustering to the deep framework. In addition, deep subspace clustering (Ji et al. 2017; Wang et al. 2020b) is also extended on the basis of traditional subspace clustering, which further expands the application (Abavisani et al. 2020; Cai et al. 2021; Wang et al. 2020c). Therefore, adopting some ideas based on multi-view clustering can facilitate our research in the process of researching MvSD.

In the past few years, some works have investigated multi-modal data, demonstrating the effectiveness of deep learning for multi-modal data fusion. Several surveys reviewed the progress of multi-view deep learning (Wang 2021; Baltrušaitis et al. 2018; Chen et al. 2020c; Summaira et al. 2021; Rahate et al. 2021; Ramachandram and Taylor 2017; Zhao et al. 2017a). Wang (2021) discussed some recent researches on deep multi-modal models from two aspects of clustering and classification, focusing on the application of generative adversarial network (GAN) in clustering and cross-modal learning. Baltrušaitis et al. (2018) investigated the latest developments in multi-modal machine learning, and it stated five challenges: representation, translation, alignment, fusion and co-learning. Chen et al. (2020c) analyzed the prevailing multi-modal network structure and existing problems, including multi-modal feature extraction and latent feature learning. Summaira et al. (2021) discussed the latest advancements and trends in multi-modal deep learning, and adopted a new fine-grained taxonomy to classify existing multi-modal networks. Rahate et al. (2021) reviewed the relevant literature in multi-modal deep learning and categorized multi-modal co-learning from multiple perspectives. These aforementioned surveys are instructive for our MvSD investigation, at the same time, some spatio temporal data (STD) analysis works provide us with application fields and current progress for reference. Wang et al. (2020d) reviewed the recent development of deep learning technology in STD and classified existing literature according to the types of STD, data mining tasks, and deep learning models. It classified the spatio temporal data into five types: event, trajectory, point reference, raster, and video. Alam et al. (2021) classified the STD analytics systems into three categories and provided definitions and related applications for STD. Moreover, it conducted investigations and discussions on existing programming languages, development tools, and data platforms. Atluri et al. (2018) summarized traditional machine learning methods in STD, and discussed related data mining problems in analyzing different STD. Mazimpaka and Timpf (2016) summarized the application of deep learning in trajectory data and traffic prediction.

In this paper, we conduct research on MvSD. Existing multi-view surveys mostly focus on the applications, neural networks and fusion methods, do not specifically consider combining multi-view and sequential data for research. In addition, the aforementioned STD

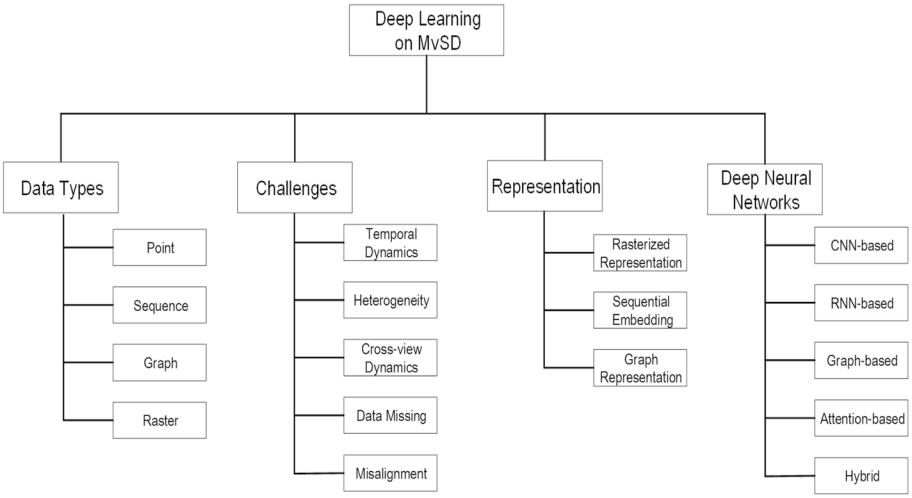


Fig. 1 Taxonomy diagram of deep learning on MvSD

researches do not consider investigating from multi-view perspective. Our contributions are as follows:

- This paper conducts research from the perspective of multi-view sequence and discusses the challenges in MvSD data.
- This survey reviews recent deep learning techniques for MvSD and categorizes MvSD into four data types. Then we organize different deep learning models for specific types of view data for representation and learning.
- This survey summarizes some application domains and emerging tasks of MvSD, and points out some potential future research directions.

The rest of this paper is organized as follows. In Sect. 2, we divide the source data that constitutes MvSD into four categories and discuss the characteristics and challenges in each of these categories. In Sect. 3, we illustrate the existing deep representation methods for various types of view data. In Sect. 4, we investigate the deep learning models for MvSD. In Sect. 5, we summarize the applications of MvSD and related tasks. Finally, we discuss the future trends and conclude the survey. The taxonomy diagram of MvSD is presented in Fig. 1.

2 Multi-view sequential data

As illustrated in Fig. 2, we give the paradigm of MvSD, which is composed of m views from different sources. These view data consists of various types, such as continuous multiple images (video clips), character description (text sentences), spatial position changes (trajectory), and these view data lengths or time steps may not be aligned with each other. Among them, each view has order constraints and is arranged according to certain rules. The sequential data of different views usually comes from various domains, which has different statistical characteristics. Therefore, it is difficult for a single view model to handle

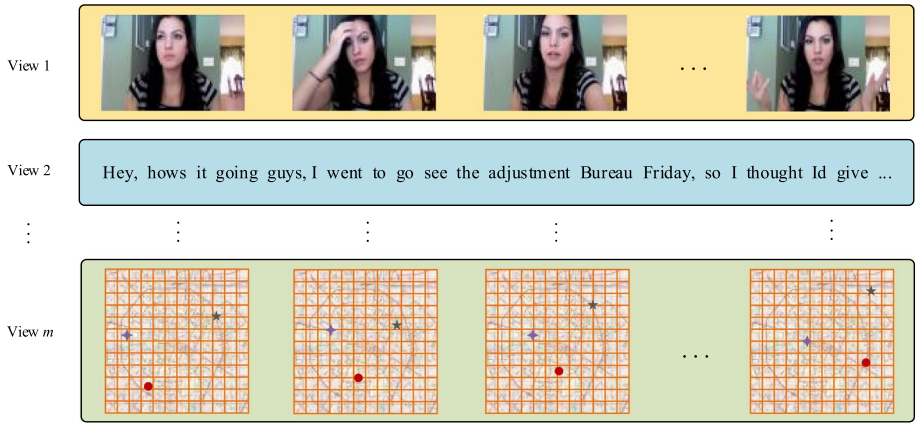


Fig. 2 Illustration of MvSD. MvSD can be composed of m views, we enumerate three types view data, including video clips, text sentences, and mobile data. Each view is arranged in a certain order (for example, chronological order, grammatical rules, etc)

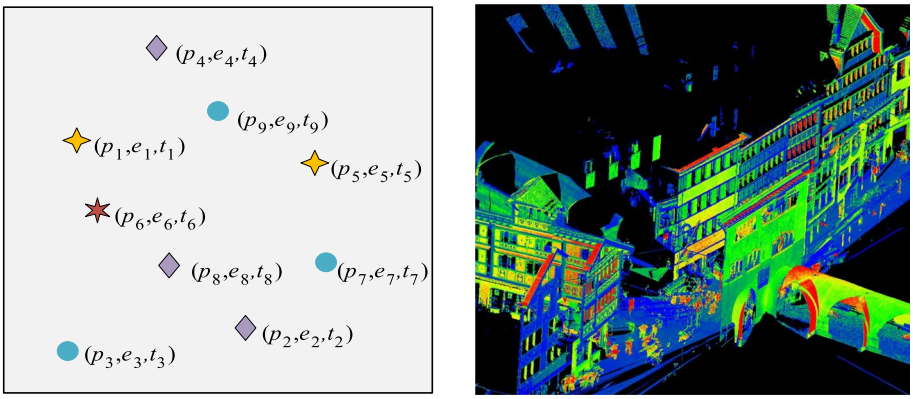
heterogeneous data. We need to formulate corresponding representation methods according to different views and select appropriate models for feature representation, extraction and fusion.

2.1 Data types

There are many types of sequential data, such as meteorological data, time-series data, gene sequences, sensor data, audio clips, etc, all of which are the research objects of MvSD. In order to facilitate subsequent work, we first introduce four data types: point, sequence, graph and raster. Each data type can be directly or indirectly converted to sequential data. Different from the classification method in Refs. Atluri et al. (2018), we generalize point data into two categories, that is, individual instances are regarded as points (e.g., event data), and the instances themselves are points (e.g., LiDAR data). In addition, we categorize trajectory data and text data into sequence data.

2.1.1 Point

Point data describes discrete points in space with specific location coordinates (e.g., geographic latitude and longitude), indicating the existence in space and attaching some additional information. A point is usually represented by a tuple (p_i, e_i, t_i), where p_i represents the position of the point, t_i represents the time when the point occurrs and e_i represents additional features (such as temperature, humidity, color, etc). Event-type data regards individual instance (e.g., a traffic incident) as point, which usually means that it occurs in a certain location and is accompanied by information such as time and event category. Figure 3a shows an example of the event data. In addition, there are some instance data themselves that are point sets, which are scanned by sensors. Point cloud data is usually represented by three-dimensional coordinates, with additional information such as reflectivity, intensity, and color, etc. Figure 3b shows an illustration of the 3D point cloud data (Hackel et al. 2017). Point data has applications in many fields, such as transportation (e.g., traffic



(a) Event data

(b) Point data from Semantic3D [Hackel et al. 2017]

Fig. 3 An illustration of event and laser data

accidents), criminology (e.g., crime incidents), social media (e.g., social event), autonomous (e.g., point cloud data), etc.

2.1.2 Sequence

Time series is a typical type of sequence data, which is a sequence obtained at consecutive and evenly spaced time points. For example, in mechanical fault diagnosis, the frequency of equipments is sampled at equal intervals. Figure 4a shows an example of audio signal. As shown in Fig. 4b, video data is viewed as a series of images arranged in chronological order. The trajectory data is also treated as a time series, which periodically records the moving position of the target. Figure 4c shows an example of the trajectory data. Time series is not the only case of sequence data, there are other cases such as text data, which need to consider the logic of language. We group trajectory data, audio, video, time series, and text into sequence data.

2.1.3 Graph

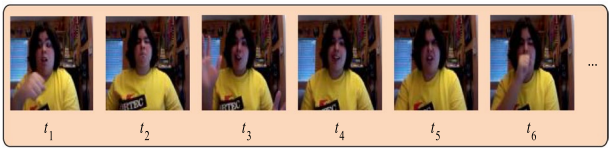
Graph data is a collection of vertices connected by a series of edges, each of which is assigned a weight. Graph data is used in many fields, including traffic networks, social networks, and recommendation systems. In a social network, each person is a vertex, and people who have a relationship with each other are connected by edges. Each edge has a direction to form a directed graph. In traffic forecasting, the traffic road networks are naturally modeled as graphs. Taking the road network as an example, where road segments are represented as edges, and nodes embedded in the spatial map represent intersections of these road segments.

2.1.4 Raster

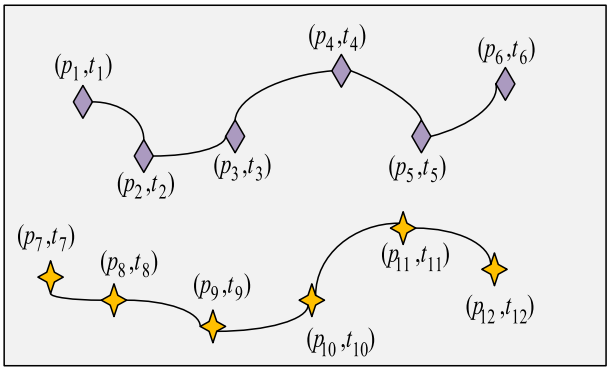
Raster data is presented in a grid of pixels, and each pixel has a value, which represents information at a specific location (color or other statistics). Figure 5a shows an example of

Fig. 4 Illustration of sequence data

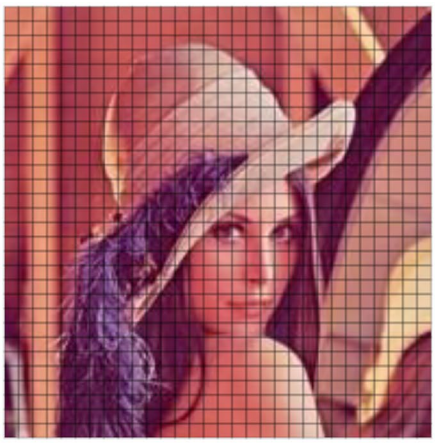
(a) Audio data



(b) Video data



(c) Trajectory data



(a) Image data

(b) Geographical raster [Zhou et al. 2020]

Fig. 5 Illustration of raster data

an image data, the position of each pixel is regarded as a fixed point, and each pixel is an observation value. In neuroscience, as a new neuroimaging method, functional magnetic resonance imaging (fMRI) is based on measuring changes in hemodynamics caused by neuronal activity. The scanned signals form raster data used to analyze brain activity. In urban big data, various fixed-position sensors collect data to form spatial map, air quality, and weather data. Figure 5b shows an example of raster traffic data (Zhou et al. 2020).

2.1.5 Converting data format

The data formats mentioned above often need to be converted into appropriate formats according to specific tasks and models. These formats are often convertible to each other. Point data is naturally converted to raster data by quantifying in each grid cell. For example, the events (e.g., traffic accidents, crimes, etc) that occur in each grid are converted into event raster data, which in turn can be converted to point data. In autonomous driving, point cloud is converted into 3d voxel grid or 2d bird's eye view (BEV) through quantization operation. Further, point data are treated as nodes in graph data. In spatial map, traffic sensors are viewed as nodes of the graph, and the distances between the sensors are used to construct an adjacency matrix. In some cases, sequence data is viewed as a series of observations in continuous time (e.g., sensor data), and the sequence data is converted to point data by sampling at equal intervals. In addition, some types of sequence data (e.g., trajectory data) are converted into raster data, and the positions of different time instants correspond to the coordinates of the grid in raster data. For some raster data (such as meteorological data), sequence data is obtained by performing continuous time statistics on the observations at each site.

2.2 Challenges

In this section, we discuss the challenges of MvSD and summarize the existing literature. We state the following five problems in MvSD: temporal dynamic, heterogeneity, cross-view dynamics, data missing, misalignment of asynchronous multi-view.

2.2.1 Temporal dynamics

For each sequential data, consequent changes are recorded in chronological order, showing the dynamics at different time slots. The data points at different times in sequential data depend on each other. If the dynamic information of the temporal granularity is ignored, the regularity of sequence change becomes difficult to model and the accuracy will decrease. For example, in the sentiment analysis task, the expressed opinions such as "*I think it's...but...*", the semantics will be inconsistent or further enhanced, which is also known as intra-modality dynamic. Information at certain moments will drive sentiment recognition. In traffic forecasting, the detected flow in the same road section is affected by human travel and often shows closeness, period and trend. In crime prediction, the factors that cause crimes may change over time. For example, there are different crime patterns on weekdays and weekends. In air quality forecasting, air quality monitoring stations record changes in the next few hours or a day. These time series show dynamic changes, and even some unexpected factors attacks may lead to sudden changes.

Early sequence modeling methods were proposed under specific tasks. For example, Prophet (Taylor and Letham 2018) was proposed by Facebook in 2017 for the company's

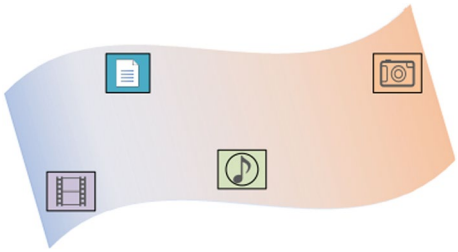


Fig. 6 Heterogeneity of MvSD. Different view data has different distribution

internal business time series. And the early air quality prediction tasks were modeled by random forest (Fawagreh et al. 2014) and inverse distance weighting (Lu and Wong 2008). Autoregression (AR) models, used to describe certain time-varying processes, such as stock forecasts (Ferenstein and Gasowski 2004), climate changes (Janjua et al. 2014). In addition, AR models and their variants are used in prognostication and health monitoring (PHM) (Barraza-Barraza et al. 2017), some variants such as autoregressive moving average (ARMA) (Pham and Yang 2010), autoregressive integrated moving average (ARIMA) (Ordóñez et al. 2019). Further, some methods based on Gaussian process (Zhao and Sun 2016a, b), Markov chain model (Sun et al. 2015), ARIMA (Chen et al. 2011), etc., have been proposed for traffic prediction.

The current schemes for modeling the temporal dynamics of multi-view sequences is to use networks based on RNNs and their derivatives. In sentiment analysis, Refs. (Zadeh et al. 2017; Verma et al. 2020; Wang et al. 2019c) employed independent LSTM to model intra-modality dynamic separately for each view sequence. To model the context of the sequence, Refs. (Hazarika et al. 2020; Xu et al. 2019) introduced bi-directional LSTM to obtain feature representations for each view. In order to obtain the temporal dependence of each weather sequence, DAQFF (Du et al. 2019) utilized bi-directional LSTM to learn long-term temporal characteristics from multivariate time series. DeepAir (Yi et al. 2018) followed the method of DeepSD (Wang et al. 2017), using RNN to embed sequence data to find similarities in different time slots.

In addition, there are some literatures that combine attention-based structures to tackle temporal dynamics. Pham et al. (2018) learnt common representations for different modalities via sequence-to-sequence (Seq2Seq) and introduced an attention mechanism to handle long-term dependencies. To address temporal duplication content in the identical view, Tian et al. (2020) adaptively aggregated useful information through self-attention. DCRNN (Li et al. 2017) captured temporal dependencies among time series through gated recurrent unit (GRU). ST-MetaNet (Pan et al. 2019) proposed a meta RNN, which uses meta-knowledge to generate GRU weights from node embeddings to model diverse temporal dependencies. Forecaster Li and Moura (2019) applied graph transformer to model long-term temporal dependencie. In order to solve the cumulative error amplification in sequence prediction, GMAN (Zheng et al. 2020) directly encoded historical inputs and generated future time steps by transform attention, thereby mitigating the error propagation problem.

2.2.2 Heterogeneity

MvSD consists of a sequence of views from multiple domains, and these views are often heterogeneous. As shown in Fig. 6, the various data mentioned in Sect. 2.1 have their own distributions. For example, image and text data are presented in different forms. Images are

usually composed of raster pixels, and the content is intuitive to humans. Whereas, textual data, which usually consists of words and symbols, follows linguistic logic and is therefore more complex than images.

In order to solve the heterogeneity of different views, models trained from specific domains are usually used to extract feature representations on corresponding view. For example, Refs. (Zadeh et al. 2017; Verma et al. 2020) extracted language, audio and visual features through three independent modality-specific LSTMs, and then explored the relationships between these modalities in the feature space. ADAIN (Cheng et al. 2018) combined feedforward neural network (FNN) and RNN, where FNN extracted static features and RNN learnt time series features. The obtained features from different views are combined for subsequent networks. stMTMV (Liu et al. 2016b) introducesd linear functions to deal with spatial and temporal features separately, and then aligned spatio-temporal views on nodes. DeepCrime (Huang et al. 2018) proposed a category-dependent encoder that encoded regions and crimes separately and finally mapped them in a common latent space.

An encoder-decoder structure is used to implement transitions between modalities to address view heterogeneity. Pham et al. (2018) translated two modalities into another joint representation via Seq2Seq model. MCTN (Pham et al. 2019) converted one modality to another via circular translation. Furthermore, for the three modalities, the representation learned between the two modalities was further transformed into the other modal, thereby forming the final joint representation. Forecaster (Li and Moura 2019) adopted the encoder-decoder architecture, taking spatial information and auxiliary information as the encoder input, and the decoder predicted the future spatial information.

The heterogeneity gap between different views is minimized in the common feature space. ARGF (Mai et al. 2020) introduced an adversarial approach to transform the distribution of the source modality into the distribution of the target modality. Inspired by domain adaptation, MISA (Hazarika et al. 2020) mapped multiple modalities into a shared subspace according to a weight-sharing encoder, and aligned these features by introducing metric distances.

2.2.3 Cross-view dynamics

MvSD has temporal dynamics within a single view sequence, while there are dynamic interactions between different view sequences. We consider cross-view dynamics into two categories, spatio-temporal correlations and semantic interactions.

Spatio-temporal correlations MvSD changes continuously in time and manifests differently in space, with spatio-temporal dynamics within a single view sequence or across views. For example, each image in a video can be viewed as a continuous change in space over a time period. For another example, in traffic forecasting, the observations of each traffic sensor are closely related to the observations of the surrounding space, and each observation value is also related to its own historical observation. There are many studies exploring spatio-temporal correlations. A conventional way is to model the local space first and then mine the temporal dynamics with recurrent networks, such as combining local convolution with recurrent networks (Yao et al. 2018; Zhou et al. 2020; Bai et al. 2019; Yu et al. 2017; Song et al. 2020; Wang et al. 2020e; Yuan et al. 2018; Chen et al. 2019; Zhang et al. 2017). Bai et al. (2019) combined graph convolutional network (GCN) with LSTM, where the local spatial correlations captured by the graph convolutional network were fed into a multi-layer LSTM to model the temporal relationships. In addition, combining attention mechanism and encoder-decoder structure is also used for spatio-temporal

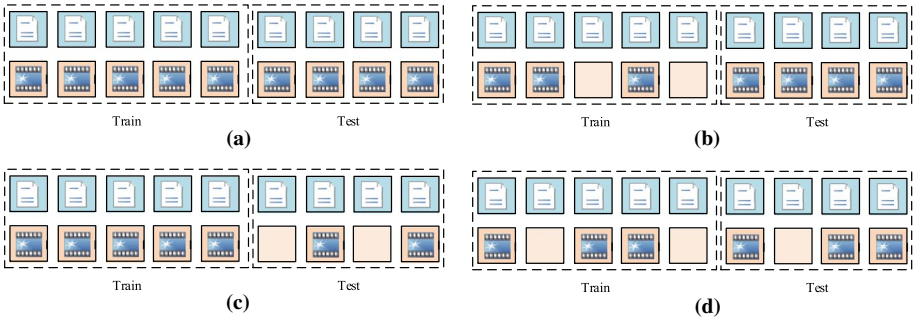


Fig. 7 Different missing types of MvSD. Taking text and video sequences as examples, (**a**) is the complete multi-view sequence. We summarize the missing types as missing in the training phase (**b**), missing in the testing phase (**c**), and missing both in the training and testing phases (**d**)

dynamic modeling (Li and Moura 2019; Shi et al. 2020; Yin et al. 2021a; Wu et al. 2020). APTN (Shi et al. 2020) used an attention-based encoder to model spatial, temporal, and periodical. The decoder introduced temporal attention to explore the dependence of the time steps. Forecaster (Li and Moura 2019) integrated the dependency graph into Transformer for forecasting spatially and temporally related data.

Semantic interactions Semantic interactions are often manifested in interactions between multiple views. For specific tasks, these views contain supplemental information that enhances specific views. Taking video sentiment analysis as an example, it usually treats language as the primary modality, as for images and audio as auxiliary modality. To model the semantic dynamics across views, memory-based methods are usually employed (Tian et al. 2020; Zadeh et al. 2018b, c; Ismail et al. 2020; He et al. 2020b). Furthermore, encoder-based methods transform multiple views to a specific view to learn a common representation (Pham et al. 2018; Xu et al. 2019; Mai et al. 2020; Hazarika et al. 2020). In addition, some literatures employ contrastive learning to achieve feature-level and semantic-level interactions (Mai et al. 2021; Liu et al. 2021; Kim et al. 2021). Mai et al. (2021) performed intra-modal/inter-modal contrastive learning and semi-contrastive learning simultaneously to ensure that the intra-modal/inter-modal dynamics are fully learned.

2.2.4 Data missing

MvSD collects data from different sources, some human factors, communication delays, and sensor failure usually cause partial or fully missing of temporal data, thus data missing is a very common phenomenon. In other words, ideally complete MvSD is rare. As shown in Fig. 7, we illustrate different types of missing data. Figure 7a is the complete multi-view sequence, each view is intact during training and testing, and different views are paired with each other. Figure 7b and Figure 7c represent missing data during training and testing phases, respectively. Figure 7d indicates that there are missing data in both training and testing phases. In this section, we introduce recent deep learning methods about data missing in MvSD.

Reconstructing missing data using autoencoders is one solution. The purpose of the autoencoder is to encode source data into latent features, and then use a decoder to decode the latent features into target domain data. DCC-CAE (Dumpala et al. 2019) combined

deep canonical correlation analysis (DCCA) with cross-modal autoencoders. DCC-CAE assumes that audio and visual modalities are available during training, but only one modality is available during testing. DCC-CAE is composed of two decoders, which input available modalities and reconstruct the corresponding missing modality representations. CPM-Nets (Zhang et al. 2020) reconstructed the complete view by constructing latent representations through structural constraints. In the unsupervised situation, CPM-Nets proposed adversarial strategies to further improve the complete representation. MCTN (Pham et al. 2019) introduced cyclic consistency loss in the process of modality translation to make the learned joint representation contain as much of all modality information as possible. MFM (Tsai et al. 2018) decomposed the multimodal representation into two factors: multimodal discriminative and modality-specific generative factors. Among them, the discriminant factors contain the shared joint features used to discriminate the task. The information contained in generative factors is unique to generate specific modalities.

Meta-learning is a learning-to-learn algorithm that learns multiple tasks on training data and processes new tasks during testing. Meta-learning enables knowledge transfer for task-agnostic few-shot learning. SMIL (Ma et al. 2017) is the first work to study the lack of data in both training and testing phases. SMIL jittered the latent feature space through Bayesian meta-learning, making single modality embeddings approximate to full modality embeddings. Meta-learning based spatio-temporal network (Yao et al. 2019a) is suitable for solving the problem of unbalanced spatial distribution of collected data, and transferring knowledge from multiple cities to target cities.

2.2.5 Misalignment of asynchronous multi-view

Sequential data from different views are usually not strictly aligned. One is that the lengths of the view sequences are not equal. For example, affected by the sampling frequency, the number of images and the number of words in the video are not equal. The other is semantic misalignment. There is no complete correspondence between images and text in video data. Each image does not correspond to each word, and one character can be associated with multiple images. Figure 8a shows the ideal condition, that is, the sequence length and semantics of multiple views are completely aligned. However, in reality, more scenarios are shown in Figure 8b and c, where Fig. 8b shows length aligned but semantics unaligned, and Fig. 8c shows both length and semantics misaligned.

Most of the existing work, such as (Zadeh et al. 2017; Wang et al. 2019c; Liang et al. 2018a), are based on the multi-view sequence alignment. Some recent work focuses on multi-view sequence misalignment (Tsai et al. 2019; Yang et al. 2020; Aytar et al. 2017; Le et al. 2018). Recently, some literatures adopt attention-based structure to achieve multi-view sequence alignment (Tsai et al. 2019; Yang et al. 2020; Le et al. 2018). Multimodal Transformer (Tsai et al. 2019) focused on the interaction between multi-view sequences of different time steps through cross-modal attention to transform one modality to another without explicitly aligning the data. Furthermore, multi-view sequence alignment is also achieved by using pretrained network. pre-trained networks. Aytar et al. (2017) trained a deep convolutional network, which used a large amount of alignment data (including language, visual and acoustic) for aligned cross-modal representation. There are other literatures that use multi-instance learning without explicit data alignment. Tian et al. (2020) formulated weakly supervised video parsing as multi-modal multi-instance learning (MMIL), and proposed MMIL pooling to aggregate multi-modal information.

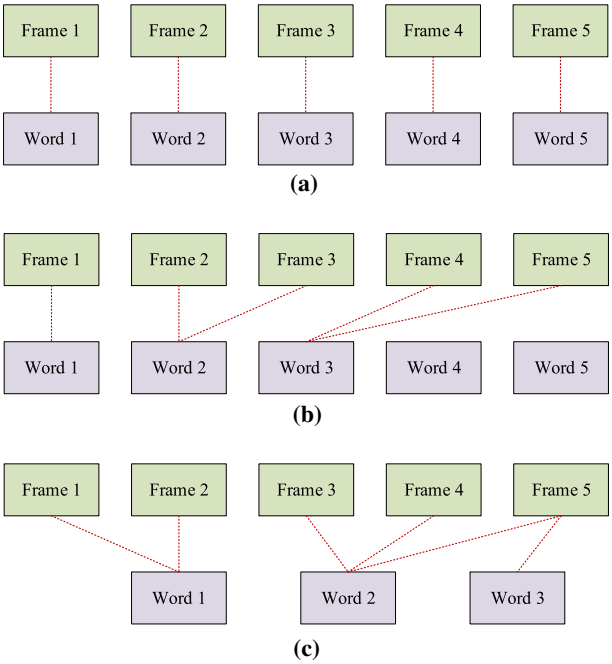


Fig. 8 Misalignment of asynchronous multi-view sequence

3 Multi-view sequential representation

In Sect. 2.1, we introduce various data forms. In order to feed these data into the network for feature learning, we need to choose appropriate methods to represent these data. In this paper, we mainly consider three representations for network input: rasterized representation, sequential embedding and graph embedding.

3.1 Rasterized representation

As shown in Fig. 9, the rasterized data quantifies the points in each grid (such as events, track points, traffic flow, meteorological data, etc.). Each cell in the raster grid is regarded as the statistics of a region. For example, in autonomous driving, points in the scene are divided into 3D rectangular grids with a given resolution (Zhou and Tuzel 2018; Yang et al. 2018; Laddha et al. 2021; Fadadu et al. 2022) to obtain voxel grids or BEVs, and then 3D CNN can be naturally applied. In traffic flow forecasting, the whole city is split to $n \times n$ grids according to latitude and longitude, and each region is indexed by rows and columns, and each grid is aggregated according to time intervals (Zhou et al. 2020; Yuan et al. 2018; Wu et al. 2020; Liao et al. 2018; Wang et al. 2020f; Zhang et al. 2019, 2021a). Through rasterization, CNN extracts the spatial features of different regions. Using recurrent networks to model raster data across multiple time slices enables analysis of dynamic relationships between different regions.

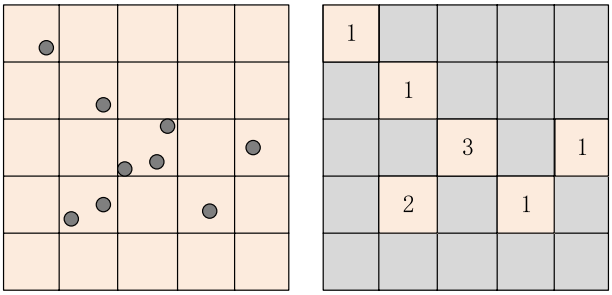


Fig. 9 Illustration of rasterized representation

3.2 Sequential embedding

For sequence data, such as text, trajectory data, and time series data, some feature transformation and feature embedding methods need to be used to process these data. For time series data, multi layer perceptron (MLP) is usually used to map it into latent vectors. Yi et al. (2018) transformed the raw features of each domain data into a low-dimensional space through embedding methods. Cheng et al. (2018) applied fully connected layers (FC) to extract features of points of interests (POIs) and meteorological features (such as weather, temperature, humidity, etc.). DAQFF (Du et al. 2019) set 1×1 convolution to transform multiple time series data. In the language model, pre-trained models are usually used to perform feature transformation on the language sequence. For example, Word2Vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014), and BERT (Devlin et al. 2018) are commonly used in natural language processing. Some literatures use pretrained 300-dimensional Glove word embeddings to encode a sequence of transcribed words into a sequence of word vectors (Zadeh et al. 2017; Wang et al. 2019c; Liu et al. 2018; Zadeh et al. 2019, 2018b). Still others use the pretrained Transformer model BERT to extract utterance level textual features (Ismail et al. 2020; Rahman et al. 2020; Yu et al. 2021b; Sun et al. 2020e).

3.3 Graph representation

Graph models a set of objects and their correlations in sentences, images, and spatial. Embedding the graph into vector representation and then seamlessly connect with GCN in subsequent processing. The graph representation reflects the association of different areas of the spatial map (Yao et al. 2018; Li and Moura 2019; Bai et al. 2019; Huang et al. 2020b; Yu et al. 2017; Wu et al. 2020). Forecaster (Li and Moura 2019) learnt the weights of the non-zero entries of the adjacency matrix by introducing a sparse linear layer, taking into account that different locations may have different dependency strengths. In the spatial map, DMVST-Net (Yao et al. 2018) employed CNN to extract local feature representation of each local region and its surrounding neighbors, and embeded the local representation into a low-dimensional representation through FC to supplement the information of the graph nodes. In spatial event prediction, Wu et al. (2020) designed an embedding component to generate an embedding vector for each event time step in each time slot. The graph representation reflects the relationship

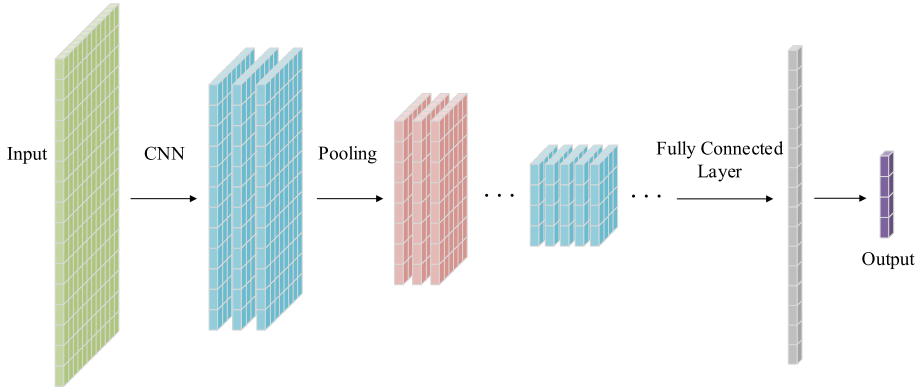


Fig. 10 Structure of CNN

between sentence and image (Wang et al. 2020a). Additionally, the graph representation also reflects the activity on social media (Islam and Goldwasser 2021).

4 Deep neural networks

In this Section, we mainly review deep network techniques commonly used to extract features of multi-view sequences. In the network, different data representations need to choose suitable network models.

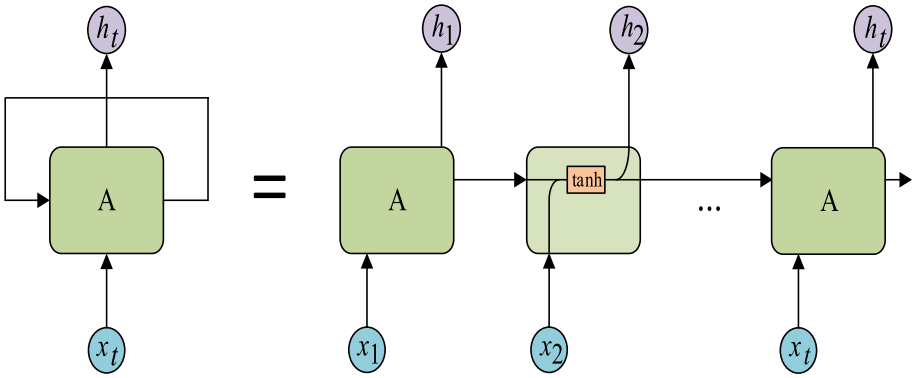
4.1 CNN-based networks

Figure 10 shows the CNN structure, which usually consists of convolutional layers, pooling layer, and fully connected layer. BatchNorm (Ioffe and Szegedy 2015) can be appended after convolutional layers. CNN has excellent performance in processing regular grid data. By stacking multiple convolutional layers, the learning from the bottom-level information to the high-level semantic features is realized in bottom-up. It moves fixed-size filters (e.g., 3×3, 5×5) on the input grid from left to right and from top to bottom. The filters performs inner product operation on corresponding position and generates high-dimensional features.

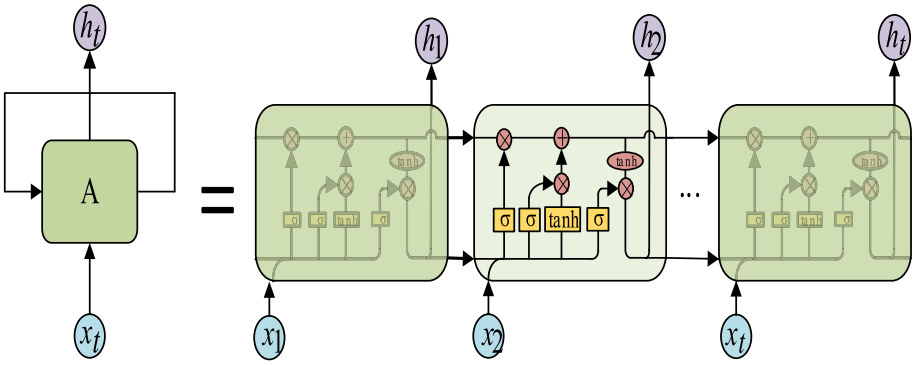
Some works try to convert traffic networks at different time intervals into raster images, and use CNN to extract spatial correlations (Zhang et al. 2017, 2019). ST-ResNet (Zhang et al. 2017) designed a CNN to capture the spatial dependency of closeness, period, and trend. MDL (Zhang et al. 2019) converted the directed graph at each time slot into a tensor representation, which was used by CNN to extract the spatial relationship, and then fully convolutional network (FCN) (Long et al. 2015) was introduced to obtain the temporal dependencies.

4.2 RNN-based networks

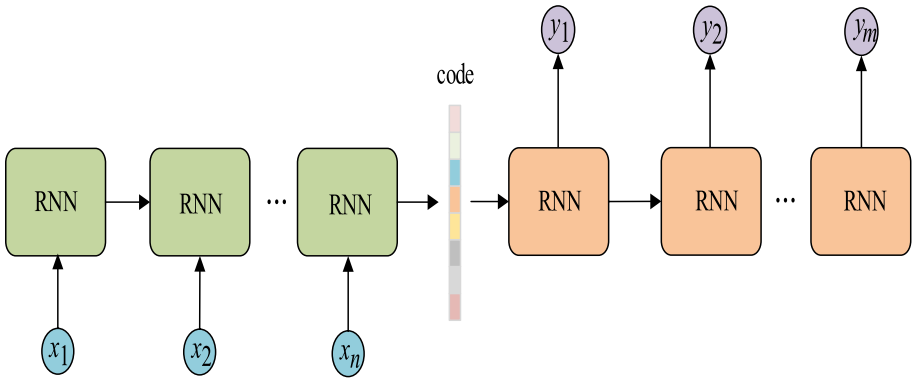
RNN and its variants are models for processing sequence data, which use the previous output state of the sequence to predict the next state, as is shown in Fig. 11a. LSTM is an extension of RNN. It has a specially customized memory unit to remember longer input



(a) RNN



(b) LSTM



(c) Seq2Seq

Fig. 11 Structures of RNN, LSTM and Seq2Seq

history information, as shown in the Fig. 11b. They are widely used in speech recognition, natural language processing and time series data analysis.

A series of RNN-based methods are used in traffic forecasting (Bai et al. 2019; Yuan et al. 2018; Wang et al. 2020f). In passenger demand prediction, three LSTMs are used to model spatiotemporal maps, external meteorological data, and temporal metadata, respectively (Bai et al. 2019). Hetero-ConvLSTM (Yuan et al. 2018) extracted spatial features through ConvLSTM (Xingjian et al. 2015), and then fed the obtained features to LSTM to model temporal dynamics. MT-ASTN (Wang et al. 2020f) modeled the temporal features of dynamic graph sequence with different scales to capture crowd flow at different scales. In the sentiment analysis task, Refs. (Zadeh et al. 2017; Verma et al. 2020; Ismail et al. 2020) utilized three independent LSTMs to construct modality embedding subnetworks for language, visual, and acoustic respectively to model intra-modality.

Figure 11c shows the structure of Seq2Seq (Sutskever et al. 2014). The Seq2Seq model is designed for sequence data and is an encoder-decoder structure, where the encoder encodes the input sequence to obtain hidden state, and the decoder generates variable-length output according to the hidden state. Pham et al. (2018) performed unsupervised learning of joint multimodal representations via Seq2Seq. Liao et al. (2018) proposed a hybrid Seq2Seq model, which integrated auxiliary information in the encoder-decoder sequence learning framework.

4.3 Graph-based networks

GCNs are often used to model non-Euclidean structural data, and GCNs are usually divided into two categories, namely spectral-based graph networks and spatial-based graph networks. The spectral-based method defines convolution operation in the Fourier domain (Kipf and Welling 2016). The spatial-based method directly applies convolution on the graph to aggregate information from neighbors. We categorize recent literatures according to the above two methods.

Spectral-based GCN introduces filters from a signal processing perspective to define graph convolution. Song et al. (2020) proposed a spatio-temporal synchronization graph convolutional network, which could effectively capture complex local spatio-temporal correlations through spatio-temporal synchronization modeling mechanism. In order to meet the requirements of medium and long-term prediction tasks, Yu et al. (2017) introduced a spatio-temporal graph convolution network, which modeled the traffic network as a graph, and used spectral convolution to extract spatial features. Geng et al. (2019) encoded the correlations of different regions to obtain multiple graphs, which were then used for correlation modeling based on ChebNet-based multi-graph convolution.

Spatial-based GCN simulates the convolution operation of traditional CNN, and graph convolution is based on the spatial relationship of nodes. Wang et al. (2020f) converted the flow Origin-Destination (OD) matrix into a semantic graph, and then performed convolution operation on the semantic graph. Li et al. (2017) treated traffic flow as a directed graph and captured spatial dependencies according to the diffusion process on the graph through a diffusion convolutional recurrent neural network. Graph attention network (GAT) (Veličković et al. 2017) is another graph neural network, which calculates the weights of neighbor nodes through an attention mechanism, without knowing the structure of the graph. Pan et al. (2019) proposed a meta-graph attention network, which used attention mechanism to capture the dynamic spatial correlation between nodes, and the attention weights were generated from meta-knowledge.

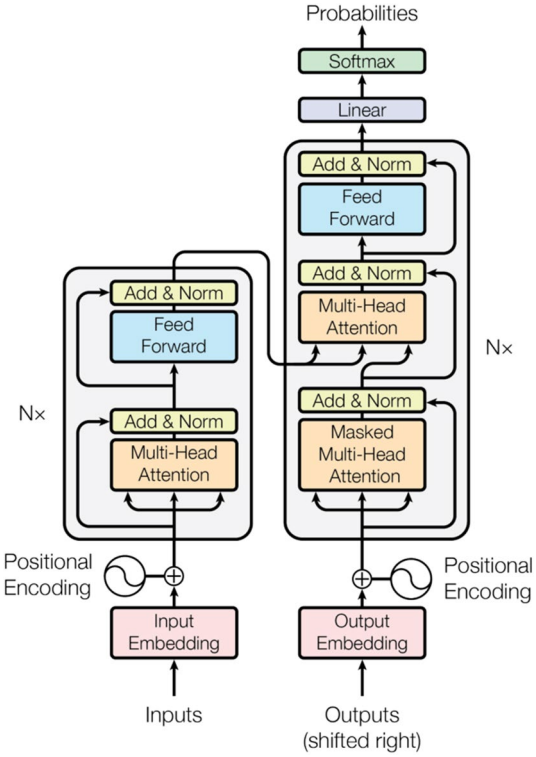


Fig. 12 Structure of Transformer (Vaswani et al. 2017)

4.4 Attention-based networks

Attention is originally proposed for natural language processing, and is now widely used in sequence-based tasks, where it models the relevant parts of the information and amplifies the most important parts. Attention can be used not only to focus on spatial dependencies, but also on spatio-temporal correlations.

Shi et al. (2020) proposed an attention mechanism to model spatial, short-term and long-term cyclic dependencies. Huang et al. (2020b) developed a graph attention network integrated with GCN into a spatial gated block to capture spatio-temporal features. GMAN (Zheng et al. 2020) introduced a transform attention layer between encoder and decoder to model the relationship between historical and future time steps. Refs. (Tian et al. 2020; Wu et al. 2021) developed hybrid attention network to jointly model temporal recurrence, co-occurrence, and asynchrony. Specifically, temporal recurrence was solved by self-attention, while co-occurrence and asynchrony were addressed by cross-modal attention. Zadeh et al. (2018b) studied a delta-memory attention network, which focused on cross-view interactions, and aggregated interactions over time with multi-view gated memory.

Transformer based on attention has been successfully applied (Tsai et al. 2019; Zadeh et al. 2019; Hasan et al. 2021; Wang et al. 2020g). The structure of the Transformer is shown in Fig 12. Tsai et al. (2019) developed a multi-modal transformer to model unaligned multi-modal language sequences, and integrated multi-modal time series from multiple pairs of cross-modal transformers. Zadeh et al. (2019) designed a multimodal

transformer layer to capture the decomposition dynamics of multi-modal data and aligned temporally asynchronous information intra-modality and inter-modality. Hasan et al. (2021) encoded multi-modal sequences separately via Transformer, and learnt to represent punchline according to the background context.

4.5 Hybrid networks

Some works are dedicated to combining multiple modules, such as combining CNN and LSTM (Xingjian et al. 2015; Zhao et al. 2019; Yao et al. 2019b). Chen et al. (2019) combined GCN with GRU, GCN was used for spatial feature extraction, and GRU was used for capturing temporal dynamics. GC-LSTM (Chen et al. 2018) introduced LSTM to model the characteristics of the dynamic graph sequence and captured the temporal characteristics of the graph sequence. Poria et al. (2017) proposed a contextual attention-based LSTM network that focuses on contextual relations, and the attention-based fusion mechanism amplified higher quality and informative modalities. Huang et al. (2020b) adopted GCN to extract spatial features and graph attention network to extract road similarity, and finally integrated these two modules through a gate structure. Pan et al. (2019) designed a combination of meta-graph attention network and meta-recurrent network, where the meta-graph attention network captured spatial correlation, and the meta-recurrent network modeled temporal correlation.

Zadeh et al. (2018c) proposed a multi-attention recurrent network, composed of long-short term hybrid memory and multi-attention block, to discover the interactions between modalities and store them in hybrid memory. Wang et al. (2019c) combined LSTM and gating mechanism, where LSTM was used to model different view sequences, and the gated modality-mixing network was used for infering nonverbal shift vector. Xu et al. (2019) adopted a combination of bi-LSTM and attention mechanism, where bi-LSTM extracted text sequences and attention mechanism learnt the alignment weights between speech and text.

4.6 Discussion

In this section, we analyze the above-mentioned deep models, as shown in Table 1. We demonstrate popular deep models, and describe them in terms of model architecture, data form, model characteristic, and application areas.

It can be seen that for sequence data processing, whether it is text, traffic flow or biological signals, RNN-based networks are general frameworks. The reason is that RNN-based networks can model the dependencies of these temporal information. For rasterized data, such as images, spatial maps, etc., CNN-based models are usually used for spatial modeling. In traffic prediction, specific nodes of some grid maps are turned into graph networks, which are suitable for GCN, spatial attention networks, etc. For multimedia data, such as text streams, audio, etc., in terms of feature representation, candidate extractors such as CNN, Transformer are available. Furthermore, it can be found that combining CNNs and recurrent networks becomes a paradigm for spatio-temporal modeling. For multi-view sequence modeling, the paradigm uses view-private recurrent networks for feature extraction, followed by cross-view interaction.

Table 1 Comparison of deep models

Models	Approach			Application	Details
	Architecture	data	characteristic		
TFN (Zadeh et al. 2017)	LSTM	Text, audio, image	RNN-based	Sentiment analysis	Using three independent LSTMs to model the temporal information of each modality separately
ST-ResNet (Zhang et al. 2017)	CNN	Rasterized map, time series	CNN-based	Traffic forecasting	This paper constructs 3 views: closeness, period and trend, and uses CNN to model spatio-temporal correlations
GSTNet (Fang et al. 2019)	GCN	Time series, graph	Graph-based	Traffic forecasting	This paper constructs a global spatio-temporal network, which uses temporal convolution to capture temporal features and a graph model to capture spatial dependencies
MLRF (Liang et al. 2018b)	LSTM	Image, audio	RNN-based	Emotion recognition	Multimodal local ranking is used to obtain relative sentiment strength, then a Bayesian ranking algorithm infers global ranking, and finally combines multimodal behavior and relative sentiment ranking for inference
ADAIN (Cheng et al. 2018)	LSTM, attention	Time series	Hybrid	Climate science	This paper models different local temporal information through LSTM, and these latent local features are finally fused through attention
DeepCrime (Huang et al. 2018)	GRU, attention	Time series	Hybrid	Crime analysis	This paper models crime dynamics through hierarchical GRU and uses attention to capture the spatial and temporal associations of different crime patterns
Pham et al. (2018)	Seq2Seq	Text, audio, image	RNN-based	Sentiment analysis	Using Seq2Seq for unsupervised joint multimodal representation learning
MISA (Hazarika et al. 2020)	LSTM	Text, audio, image	RNN-based	Sentiment analysis	Using 3 independent LSTMs to extract the features of text, audio and image respectively. Specifically, to model intra-modal and inter-modal relationships, MISA performs distance metrics in two subspaces

Table 1 (continued)

Models	Approach			Application	Details
	Architecture	data	characteristic		
HybridAtt (Yuan et al. 2019)	CNN, Attention	Time series	Hybrid	Sleep staging	1D CNN is used to process multivariate polysomnography records, and the designed hybrid attention mechanism further fuses multi-view features
MetaST (Yao et al. 2019a)	CNN, LSTM	Rasterized map	Hybrid	Water quality prediction	A meta-learning paradigm spatiotemporal network is designed to solve the data imbalance problem through transfer learning
Yuan et al. (2018)	ConvLSTM	Rasterized map	Hybrid	Accident prediction	A Hetero-ConvLSTM is proposed to handle both spatial heterogeneity and temporal autocorrelation
GraphSleepNet (Jia et al. 2020)	GCN, attention	Rasterized map	Hybrid	Sleep staging	Based on spatiotemporal graph convolution network, graph convolution extracts spatial features and temporal convolution is used to capture transition rules between sleep stages
Hasan et al. (2021)	Transformer	Text, audio, image	Attention-based	Humor detection	Language-, acoustic-, visual-, and humor-centric features are extracted separately using Transformers, followed by cross-attention layers for information interaction
Tian et al. (2020)	Attention	Image, audio	Attention-based	Video parsing	A hybrid attention network is proposed to simultaneously explore unimodal and cross-modal temporal context. A multimodal multiple instance learning (MMIL) pooling method is developed to adaptively explore useful audio and video content

5 Applications

In this section, we summarize related work in different domains, including intelligent transportation, crime analysis, sentiment analysis, climate science and health care. We discuss these application areas separately and provide an overview of recent related techniques. In Table 2, we enumerate the application areas along with the aforementioned models.

5.1 Intelligent transportation

With the development of mobile communication and human daily travel, a large amount of traffic data is generated, which usually includes traffic volume, speed, accidents, trajectories, spatial maps, and road networks, etc. Traffic data mining and analysis has become an urgent problem to be solved. Traffic forecasting plays an important role in smart cities, providing constructive guidance for urban planning, intelligent management and public safety, thereby promoting urban construction and avoiding waste of resources. Table 3 summarizes the performance of spatio-temporal models on several publicly available benchmark datasets. In TaxiBJ[1] and NYC Bike[2], we report root mean square error (RMSE) in terms of both flow and demand. For the two types of graph structure data, PeMSD4[3] and MeTR-LA[4], we mainly summarize RMSE and mean absolute error (MAE).

As mentioned in Sect. 2.1, traffic data is typically represented as trajectories, events, and raster data. Among them, data such as trajectories and events are usually converted to raster data for processing, and then this type of data is consumed by networks such as CNN (Lv et al. 2019; Chen et al. 2020; Guo et al. 2019a; Sun et al. 2020a). The traffic forecasting of a single road segment can be regarded as sequence data, which is fed to RNN or LSTM (Huang et al. 2014; Yang et al. 2016). The road network is represented using a graph, and these road network associations are modeled by GCN (Yu et al. 2017; Geng et al. 2019; Fang et al. 2019; Sun et al. 2020b; Bai et al. 2021). In addition to the diversity of data forms, complex spatial-temporal correlations hinder traffic prediction. Among them, the spatial correlation is mainly reflected in different regions and different road segments. Traffic in adjacent areas is spatially causal, i.e. flows from one area to another. The temporal correlation is reflected in the fact that the flow of the area is affected by different time intervals, such as short-term or long-term changes. Therefore, spatial-temporal based models are used to deal with this problem (Bai et al. 2019; Zhao et al. 2019; Guo et al. 2019b).

Further, traffic data is affected by external factors (such as weather, accidents), and the data comes in various forms. Some literatures (Zhou et al. 2020; Wang et al. 2020f; Zhang et al. 2019) model the temporal dynamics into three views: closeness, period, and trend. In traffic flow forecasting, Guo et al. (2019b) combined spatio-temporal attention

[1] TaxiBJ is the taxicab GPS data and meteorology data in Beijing from four time intervals: 1st Jul. 2013–30th Otc. 2013, 1st Mar. 2014–30th Jun. 2014, 1st Mar. 2015–30th Jun. 2015, 1st Nov. 2015–10th Apr. 2016.

[2] The bike trajectories are collected from NYC CitiBike system. There are about 13000 bikes and 800 stations in total.

[3] PeMSD4 describes the San Francisco Bay Area, and contains 3848 sensors on 29 roads dated from 1/1/2018 until 2/28/2018, 59 days in total.

[4] MeTR-LA records four months of statistics on traffic speed, ranging from 3/1/2012 to 6/30/2012, including 207 sensors on the highways of Los Angeles County.

Table 2 Deep models for different applications of MvSD

	Transportation	Health care	Crime analysis	Sentiment analysis	Climate science
CNN-based	Zhou et al. (2020), Zhang et al. (2017, 2019), Lv et al. (2019), Chen et al. (2020), Guo et al. (2019a), Sun et al. (2020a)	Akman et al. (2021), Jia et al. (2021b), Nessiem et al. (2021), Coppock et al. (2021), Piriyajitakonkij et al. (2020)	Wang et al. (2019a), Okawa et al. (2019), Salama et al. (2021)		Yi et al. (2018)
RNN-based	Yuan et al. (2018), Yao et al. (2019a), Wang et al. (2020f), Huang et al. (2014), Yang et al. (2016)	Le et al. (2018), Phan et al. (2019, 2021), Olesen et al. (2021)	Huang et al. (2018)	Zadeh et al. (2017), Hazarika et al. (2020), Pham et al. (2019), Liu et al. (2018), Sun et al. 2020e; Mai et al. 2019)	Du et al. (2019, 2021), Cheng et al. (2018), Veiga et al. (2021)
Graph-based	Pan et al. (2019), Song et al. (2020), Geng et al. (2019), Fang et al. (2019), Sun et al. (2020b), Bai et al. (2021), Chen et al. (2020), Guo et al. (2020)	Jia et al. (2020, 2021a)	Zhang et al. (2021b), Wang et al. (2021), Xia et al. (2021)	Mai et al. (2020), Wang et al. 2022)	Liu et al. (2016b), Chen et al. (2021c)
Attention-based	Zheng et al. (2020), Shi et al. (2020), Huang et al. (2020b), Guo et al. (2019b)	Phan et al. (2022)	Rayhan and Hashem (2020)	Tian et al. (2020), Zadeh et al. (2018b), Wu et al. (2021), Hasan et al. (2021), Brousmiche et al. (2021), Gu et al. (2018), Akhtar et al. (2019)	
Hybrid	Yao et al. (2018, 2019b), Li et al. (2017), Bai et al. (2019), Yu et al. (2017), Zadeh et al. (2019), Zhao et al. (2019), Guo et al. (2019b)	Yuan et al. (2019), Supratak et al. (2017)	Stec and Klabjan (2018), Ertugrul et al. (2019)	Xu et al. (2019), Zadeh et al. (2018c), Tsai et al. (2019), Zadeh et al. (2019), Bie and Yang (2021), Chen et al. (2017), Han et al. (2021a)	Cheng et al. (2018), Yao et al. (2019a), Zhong et al. (2020), Han et al. (2021c), Sasaki et al. (2021), Ouyang et al. (2021), Civitarese et al. (2021), Lin et al. (2020), Lu and Li (2020)

Table 3 Various traffic prediction models performance on TaxiBJ, NYC Bike, PeMSD4, and MeTR-LA

Models	TaxiBJ		NYC Bike		Models	PeMSD4		MeTR-LA	
	Flow (RMSE)	Demand (RMSE)	Flow (RMSE)	Demand (RMSE)		RMSE	MAE	RMSE	MAE
HA	57.69	40.439	–	8.541	HA	54.14	36.76	7.8	4.16
ARIMA	22.78	–	10.07	–	ARIMA	58.05	35.19	10.45	5.15
VAR	22.88	–	9.92	–	VAR	51.62	33.63	9.13	5.41
DCRNN (Li et al. 2017)	–	20.569	–	5.215	DCRNN (Li et al. 2017)	37.12	24.55	6.383	3.125
STGCN (Yu et al. 2017)	–	19.101	–	4.759	STGCN (Yu et al. 2017)	38.16	27.02	7.24	3.47
DeepST (Zhang et al. 2016)	18.18	–	7.43	–	TGCN (Zhao et al. 2019)	–	–	8.152	4.319
ST-ResNet (Zhang et al. 2017)	16.69	17.649	6.33	6.159	AGCRN (Bai et al. 2020)	32.26	19.83	6.535	3.209
ConvLSTM (Xingjian et al. 2015)	–	18.788	–	4.745	DGCN (Guo et al. 2020)	31.61	19.6	–	–
DMVST-Net (Yao et al. 2018)	–	18.206	–	4.766	ASTGCN (Guo et al. 2019b)	32.82	21.8	6.78	3.841

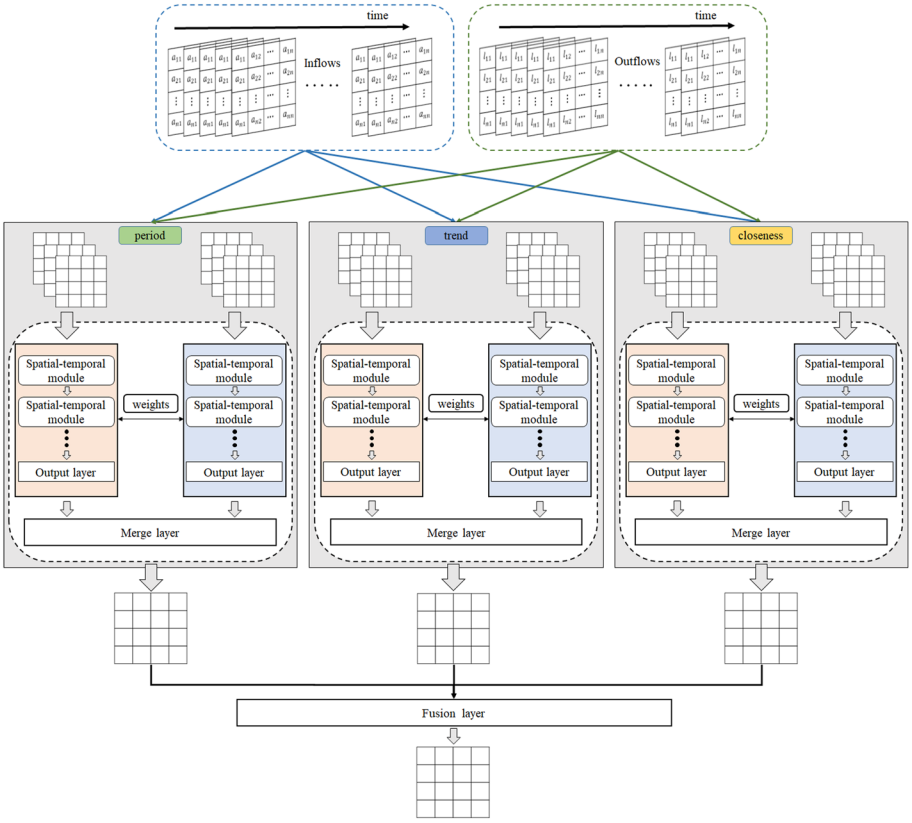


Fig. 13 Structure of DFFSTM (Zhou et al. 2020)

and spatio-temporal convolution to construct three basic components to model the recent, daily-periodic and weekly-periodic of traffic flow, respectively. In demand forecasting, Bai et al. (2019) took the structured city as one view, external meteorological data and time meta as the other two views. For the spatial view, the spatial features of each time slot were extracted by GCN, and then the spatio-temporal correlations were captured by LSTM. For the other two views, these factors were used to model passenger demand. Yao et al. (2018) built three views: spatial view, temporal view, and semantic view. Geng et al. (2019) studied multiple graphs to model the relationship between different regions, including neighborhood, functional similarity, and connectivity.

Example 1 Zhou et al. (2020) proposed a deep flexible structured spatial–temporal model (DFSSTM) for taxi capacity prediction. The structure of DFSSTM is shown in Fig. 13. First, the traffic data was rasterized into a $n \times n$ grid, and the flow relationships of vehicles were characterized by inflows and outflows. Due to the short-term and long-term dependencies of traffic data, DFSSTM divided temporal dependencies into three views: period, trend and closeness. Subsequently, DFSSTM tailored a siamese spatio-temporal network (SSTN), which took both inflows and outflows as inputs, to model spatio-temporal dependencies. Three SSTNs were used to model three temporal dynamics (period, trend

Table 4 Performance of sleep staging models on SleepEDF-20, MASS, and SHHS

Models	SleepEDF-20			MASS			SHHS		
	Acc	κ	MF1	Acc	κ	MF1	Acc	κ	MF1
DeepSleepNet (Supratak et al. 2017)	82.0	0.760	76.9	86.4	0.805	82.2	–	–	–
FCNN+RNN (Phan et al. 2021)	83.5	0.775	77.7	86.4	0.806	82.1	88.1	0.832	80.9
TinySleepNet (Supratak and Guo 2020)	85.4	0.800	80.5	83.1	0.77	78.1	–	–	–
RobustSleepNet (Guillot and Thorey 2021)	–	–	81.7	–	–	82.5	–	–	80.0
SleepTransformer (Phan et al. 2022)	–	–	–	–	–	–	87.7	0.828	80.1
SeqSleepNet (Phan et al. 2019)	86.0	0.809	79.7	87.0	0.815	0.833	88.4	0.838	80.1
XsleepNet (Phan et al. 2021)	86.4	0.813	80.9	87.6	0.823	83.8	89.1	0.847	82.3
SalientSleepNet (Jia et al. 2021b)	87.5	–	83.0	–	–	–	–	–	–

and closeness). Finally, DFSSTM designed a fusion layer that automatically adjusted the weights to integrate different views.

5.2 Health care

In the medical system, clinicians diagnose patients through comprehensive consideration of multiple factors (for example, previous medical history, various biological indicators, patient physique, etc). This process is complicated and time-consuming. In order to better assist clinicians in diagnosis, many works apply deep learning technology in the field of medical intelligence (Yuan et al. 2019; Jia et al. 2020; Akman et al. 2021; Phan et al. 2022; Olesen et al. 2021; Piriyajitakonkij et al. 2020; Feng et al. 2021; Torres et al. 2016). The intelligent system discovers disease patterns by learning historical data and various clinical indicators, supplemented by experts knowledge. Table 4 summarizes the performance of the automatic sleep staging models on the SleepEDF-20[5], MASS[6], and SHHS[7] datasets. We report overall accuracy (Acc), Cohen's kappa (κ), and macro F1-score (MF1).

Clinical data usually comes from multiple views and is heterogeneous (for example, verbal description, medical imaging, polysomnography, etc.). To handle asynchronous view sequences, Le et al. (2018) investigated memory technology to establish cross-view interactions and dependencies. It encoded the input view separately through two encoders and saved to two external memories. In sleep monitoring, Phan et al. (2021) took the original signal and time-frequency information as input. In order to solve the over-fitting rate between different views, it dynamically adjusted the learning steps between different modalities, and derived weights based on specific views for fusion different views. Jia et al. (2021b) proposed a temporal fully convolutional network based on U^2-Net (Qin et al. 2020) for multimodal salient waves detection, which converted the time series classification problem into a salient detection problem. At the same time, the multi-scale features of the sleep

[5] SleepEDF-20 is the Sleep Cassette subset of the Sleep-EDF Expanded dataset Kemp et al. (2000), consisting of 20 subjects (10 males and 10 females) aged 25–34.

[6] MASS is pooled from different hospital-based sleep laboratories, consisting of whole-night recordings from 200 subjects (97 males and 103 females) aged 18–76.

[7] The SHHS database Zhang et al. (2018) has two rounds of PSG records, namely Visit 1 and Visit 2. The former, consisting of 5791 subjects aged 39–90, was employed in this work.

stage are captured by the multi-scale extraction module. In recent years, the coronavirus (COVID-19) has spread worldwide, causing a large number of human casualties and economic losses. Some works use coughing and breathing audio to determine whether COVID is positive or negative (Akman et al. 2021; Nessiem et al. 2021; Coppock et al. 2021).

5.3 Crime analysis

Crime prediction plays a vital role in crime prevention. Recent works (Huang et al. 2018; Okawa et al. 2019; Stec and Klabjan 2018; Vomfell et al. 2018) can dig out the spatial-temporal patterns and trends of crimes through historical data combined with information such as crime incidents, time and location. This contributes to the early warning, allowing police to conduct inspections in high-risk regions, reducing the impact on society. However, unlike traffic data, the distribution of crime incidents in spatial and temporal is sparse, and there are fewer spatial-temporal associations between different crime incidents.

In order to solve the high heterogeneity of crime spatial distribution, the city is usually converted to raster data to extract spatial attributes, and then a recurrent network is used to capture the temporal dynamics. Wang et al. (2019a) selected crime incidents in a specific area, and divided the area into 16×16 raster images. Motivated by ST-ResNet (Zhang et al. 2017); Wang et al. 2019a) constructed three views to extract nearby, periodic, and trend features separately. To model the association between crime and different regions, DeepCrime (Huang et al. 2018) proposed a category-interactive encoder, which embeds information such as spatial and event categories into latent vectors for representation. In addition, DeepCrime (Huang et al. 2018) adopted three GRUs to separately encode crime sequences, abnormal sequences, and interdependent sequences to model the temporal dynamics. In order to make the model interpretable, Rayhan and Hashem (2020) considered an attention-based spatio-temporal network, which captured the dynamic spatio-temporal correlation of crimes based on past criminal events, external features and recurring trends. Specifically, two GAT variants were used to embed spatial hierarchical information and specific category features respectively. CASTNet (Ertugrul et al. 2019) designed a community-attentive spatio-temporal model to capture the spatio-temporal pattern of criminal events, which was used to predict opioid overdose. Specifically, CASTNet (Ertugrul et al. 2019) extracted opioid overdose at different locations through a multi-head attention network, and introduced hierarchical attention to allow interpretation of the contribution of features from different communities to local incident prediction. Table 5 summarizes the performance of the deep crime models on the New York City[8] and Chicago[9] datasets. Note that most of these models are derived from spatio-temporal models.

5.4 Sentiment analysis

The opinions expressed by humans in daily communication are usually complex and multimodal, and it is of great significance for computer intelligence to understand these data. Sentiment analysis is also an important branch of future human-computer interaction.

In multi-modal sentiment analysis, these multi-view data are usually represented by vision, acoustic, and language. There are two challenges to be overcome in this task:

[8] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
[9] https://www.kaggle.com/datasets/chicago/chicago-crime

Table 5 Summary of deep crime models performance on New York City, and Chicago

Models	New York City						Chicago					
	Categories				Overall		Categories				Overall	
	Burglary	Larceny	Robbery	Assault	MAE	MAPE	Theft	Battery	Assault	Damage	MAE	MAPE
SVM (Chang and Lin 2011)	0.449	0.565	0.553	0.426	1.086	0.683	0.654	0.643	0.507	0.631	0.990	0.586
ARIMA (Pham et al. 2019)	0.405	0.488	0.532	0.411	0.879	0.551	0.405	0.488	0.532	0.411	0.860	0.506
ST-ResNet (Zhang et al. 2017)	0.487	0.602	0.564	0.532	0.996	0.580	0.551	0.676	0.594	0.682	0.909	0.450
STGCN (Yu et al. 2017)	0.547	0.618	0.614	0.510	0.620	0.413	0.686	0.689	0.629	0.712	0.682	0.462
DeepCrime (Huang et al. 2018)	0.468	0.647	0.636	0.536	0.623	0.415	0.624	0.705	0.641	0.731	0.677	0.455
DCRNN (Li et al. 2017)	0.533	0.624	0.610	0.519	0.654	0.451	0.683	0.647	0.598	0.679	0.681	0.471
ST-MetaNet (Pan et al. 2019)	0.538	0.553	0.535	0.495	0.624	0.418	0.698	0.619	0.656	0.726	0.758	0.480
STDN (Yao et al. 2019b)	0.526	0.542	0.588	0.471	0.898	0.471	0.629	0.662	0.586	0.634	0.953	0.505
UrbanFM (Liang et al. 2019)	0.588	0.466	0.639	0.560	0.810	0.574	0.689	0.670	0.582	0.646	0.886	0.613
ST-SHN (Xia et al. 2021)	0.617	0.650	0.672	0.595	0.583	0.357	0.715	0.738	0.674	0.729	0.643	0.400

intra-modality dynamics and inter-modality dynamics. Sun et al. (2020c) proposed a multi-view CRF model, which captured the correlation between features within a single view and considered the relationships between different views. Refs. (Zadeh et al. 2017; Liu et al. 2018) modeled the dynamics of specific modality sequences through three LSTMs and captured the interactions between modalities through a three-fold Cartesian product. Zadeh et al. (2018b) proposed a memory enhancement network, which modeled the interaction of multiple view sequences over time by introducing gated memory. Since different modalities are heterogeneous, MISA (Hazarika et al. 2020) mapped different modalities to a common subspace to learn shared feature representations. Mai et al. (2020) designed an adversarial encoder-decoder structure to embed different modalities into a common space and learn invariant representations of modalities. In order to deal with unaligned view sequences, Xu et al. (2019) introduced an attention mechanism to align speech and text, and realized the integration of speech and text at the word level. Table 6 summarizes the model performance on MOSI[10], MOSEI[11], and SIMS[12], and the evaluation metrics follow previous work Rahman et al. (2020).

In audio-visual event recognition, Brousmiche et al. (2021) studied a multi-level attention fusion network, which could dynamically integrate visual and audio information for event recognition. Tian et al. (2020) presented a new task, named audio-visual video parsing (AVVP), which detected video events (labels them as audible, visible, or audible and visible) and located the duration via a weakly supervised approach. In this task, events may occur repeatedly in different views, so there are three challenges: unimodal-modal temporal recurrence, cross-modal co-occurrence, and cross-modal asynchronous. Tian et al. (2020) simultaneously employed a hybrid attention network to solve these problems, that is, using a self-attention mechanism to model unimodal-modal temporal recurrence, and using a cross-modal attention mechanism to simultaneously deal with cross-modal co-occurrence and cross-modal asynchronous.

5.5 Climate science

Weather data is usually collected by various sensor devices, including temperature, humidity, wind speed, pressure, air quality, etc. By studying the interrelationship of these meteorological data, it is helpful for human to further understand the earth's environment and prevent natural disasters in advance.

Recently, some deep learning methods have been successfully applied to air quality prediction (Yi et al. 2018; Zhong et al. 2020; Sasaki et al. 2021; Ouyang et al. 2021; Lin et al. 2020). Refs. (Cheng et al. 2018; Du et al. 2021; Han et al. 2021c) adopted FC to extract local spatial features and LSTM to capture temporal dynamics. Yi et al. (2018) proposed a patial transformation component to aggregate spatial monitoring data of different scales. In addition, in order to model dynamic changes between cross-modal data, these features are fused in a distributed manner. Zhong et al. (2020) introduced reinforcement learning to predict air quality. The model mainly consisted of two components: site selector and air quality regressor. Among them, the site selector adaptively selected the relevant sites, and

[10] CMU-MOSI Zadeh et al. (2016) dataset is one of the most popular benchmark datasets for MSA.

[11] Compared to CMU-MOSI, CMU-MOSEI Zadeh et al. (2018a) dataset extends its data with more utterances, more samples, speakers, and topics.

[12] SIMS Yu et al. (2020b) dataset is a Chinese MSA benchmark with fine-grained modal annotations.

Table 6 Performance of multimodal sentiment analysis models on MOSI, MOSEI, and SIMS

Models	MOSI				MOSEI				SIMS			
	Acc-2	F1	MAE	Corr	Acc-2	F1	MAE	Corr	Acc-2	F1	MAE	Corr
TFN (Zadeh et al. 2017)	–/80.8	–/80.7	0.901	0.698	–/82.5	–/82.1	0.593	0.700	78.38	78.62	0.432	0.591
LMF (Liu et al. 2018)	–/82.5	–/82.4	0.917	0.695	–/82.0	–/82.1	0.623	0.667	77.77	77.88	0.441	0.576
MFN (Zadeh et al. 2018b)	77.4/–	77.3/–	0.965	0.632	76.0/–	76.0/–	–	–	77.90	77.88	0.435	0.582
RAVEN (Wang et al. 2019c)	78.0/–	76.6/–	0.915	0.691	79.1/–	79.5/–	0.614	0.662	–	–	–	–
MFM (Tsai et al. 2018)	–/81.7	–/81.6	0.877	0.706	–/84.4	–/84.3	0.568	0.717	–	–	–	–
Mult (Tsai et al. 2019)	81.5/84.1	80.6/83.9	0.861	0.711	–/82.5	–/82.3	0.580	0.703	78.56	79.66	0.453	0.564
MAG-BERT (Rahman et al. 2020)	84.2/86.1	84.1/86.0	0.712	0.796	84.7/–	84.5/–	–	–	–	–	–	–
MISA (Hazarika et al. 2020)	81.8/83.4	81.7/83.6	0.783	0.761	83.6/85.5	83.8/85.3	0.555	0.756	–	–	–	–
Self-MM (Yu et al. 2021b)	84.0/85.98	84.42/85.95	0.713	0.798	82.81/85.17	82.53/85.3	0.530	0.765	80.04	80.44	0.425	0.595
MMIM (Han et al. 2021b)	84.14/86.06	84.00/85.98	0.700	0.800	82.24/85.97	82.66/85.94	0.526	0.772	–	–	–	–

Table 7 Air quality models performance on Beijing and LongDon

Datasets	Models	Overall		PM2.5	
		RMSE	MAE	RMSE	MAE
Beijing	RF (Fawagreh et al. 2014)	26.68	15.59	–	–
	IDW (Lu and Wong 2008)	48.09	34.79	28.41	18.35
	KNN	37.99	23.94	18.24	10.58
	ADAIN (Cheng et al. 2018)	29.39	18.83	15.06	8.28
	ANCL (Patel et al. 2022)	24.28	15.23	–	–
	MCAM (Han et al. 2021c)	–	–	12.57	6.83
LongDon	RF (Fawagreh et al. 2014)	4.69	3.06	–	–
	IDW (Lu and Wong 2008)	8.01	5.50	–	–
	KNN	4.75	3.20	5.77	3.90
	ADAIN (Cheng et al. 2018)	4.78	3.36	3.11	2.02
	ANCL (Patel et al. 2022)	4.65	3.20	–	–
	MCAM (Han et al. 2021c)	–	–	2.81	1.78

the quality regressor received the selected sites for air quality estimation. In water quality prediction, Liu et al. (2016b) studied a multi-task multi-view method to fuse the data from different domains to predict the water quality of a site. In extreme weather forecasting, Civitarese et al. (2021) proposes a temporal fusion transformer, which uses multiple variables (such as static, historical, and future) as input. In order to solve the imbalance in the spatial distribution of the collected data, Yao et al. (2019a) constructed a spatio-temporal network based on meta-learning, which transfered knowledge from multiple cities to help the target city make spatio-temporal predictions. Table 7 summarizes the performance of the deep air models on Beijing[13] and Longdon[14] dataset, we focus on the overall performance and the performance on PM2.5.

Example 2 Du et al. (2019) proposed a deep air quality forecasting framework (DAQFF) for PM2.5 prediction. Figure 14 illustrates the network architecture of DAQFF. First, DAQFF customized multiple 1D convolutional neural networks for multivariate time series to model local features. Different from some spatio-temporal models, DAQFF concatenated the temporal features of multiple stations, which could simultaneously capture local features and spatial relationships across stations. Subsequently, to model long-term dependencies, DAQFF introduced bi-LSTM to extract long-term temporal dependencies. Finally, the obtained shared features were concatenated and fused for prediction.

[13] This data collected in Beijing, China. The air quality data are recorded at 36 air quality stations every hour, from 2014/05/01 to 2015/04/30.

[14] https://data.london.gov.uk/air-quality/

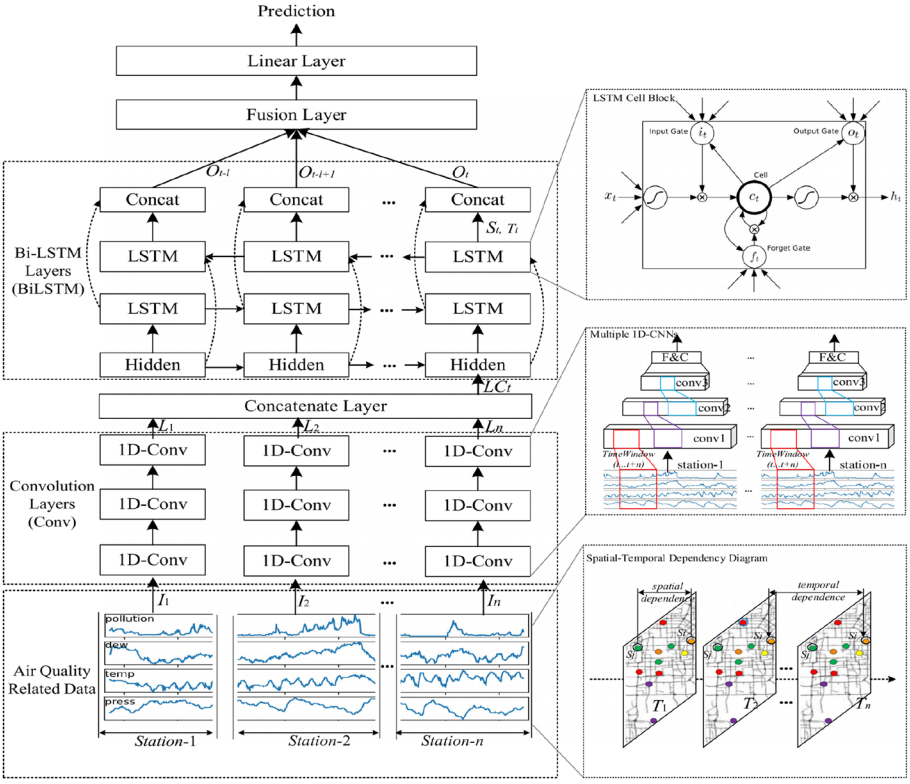


Fig. 14 Structure of DAQFF (Du et al. 2019)

6 Future directions

In this Section, we enumerate some challenging deep learning techniques in MvSD in recent years, and point out potential future research directions.

6.1 Interpretable model research

Despite the impressive achievements of deep learning, the working principle of the model is in a black box, and the decision-making is difficult to establish a reasonable basis. Therefore, interpretability models have become a research hotspot, which builds trust with users to understand why decisions are made in different domains such as medical diagnosis, autonomous driving, or recommender systems.

Multi-view data tends to introduce bias because these views are heterogeneous and each view may have deviations, resulting in model error accumulation and amplification. Some works focus on feature-level interpretability (Rayhan and Hashem 2020; Ertugrul et al. 2019; Khanehzar et al. 2021), which achieves global interpretation by modeling local feature relationships. Meanwhile, attention is given to providing reliable explanations behind the predictions (Jia et al. 2021a; Zheng et al. 2021; Ma et al. 2018; Agyemang et al. 2020). Furthermore, it is also an approach to achieve interpretability by designing

network architectures that conforms to human cognition (Choe et al. 2021). At present, there are no mature techniques and standards to estimate the performance of interpretability. Thus, it is impossible to compare the pros and cons of interpretable methods. In the future, interpretability will be further refined to guide the agent's behavior or multi-view fusion decision-making.

6.2 Multi-modal architecture research

As the network structure becomes more and more complex, the cost and time of manually designing the network will be unbearable. Especially for MvSD, in the feature extraction stage, a feature extractor needs to be designed for a single view and there are many alternative network structures for different views. In the multi-view feature fusion stage, we need to consider aggregating multiple views strategies. The automated process of neural architecture search (NAS) can speed up research on MvSD.

Auto-MVCN (Li et al. 2020) tailored a multi-view architecture for 3D shape recognition, which explored correlations between view features by automatically searching for fused cells. In electronic health records, MUFASA (Xu et al. 2021) simultaneously searched modality-specific networks and feature fusion strategies. BM-NAS (Yin et al. 2021b) designed a bilevel search scheme, BM-NAS selected feature pairs from pre-trained unimodal and searched a feature fusion strategy. To make the model applicable to various multimodal tasks, MMnas (Yu et al. 2020a) defined a general network search framework to design task-specific heads for different tasks on a unified backbone. MFAS (Pérez-Rúa et al. 2019) found a reasonable network structure for multimodal fusion by constraining the search space and employing sequential model-based exploration methods. Although many recent works try to design fusion strategies, how to perform fusion of multi-view sequences is not well studied and needs more research in the future. Through multi-modal network architecture search, better models and fusion methods are obtained..

6.3 Data annotations

Deep learning benefits from massive amounts of data, however large-scale data annotation brings prohibitive costs, which becomes more severe when annotating MvSD. Therefore, some techniques based on unsupervised, semi-supervised, etc., are introduced to facilitate MvSD research.

Unsupervised learning uses unlabeled data. In multi-view representation learning, DUA-Nets (Geng et al. 2021) combined inverse networks through unsupervised learning to automatically evaluate the quality of different views. Through unsupervised training, contrastive learning has achieved great success in the computer vision domain (He et al. 2020a). In multi-modal sentiment analysis, Mai et al. (2021) performed intra-modal/inter-modal contrastive learning and semi-contrastive learning simultaneously to ensure that the intra-modal/inter-modal dynamics are fully learned. In semi-supervised learning, a small amount of labeled data is combined with a large amount of unlabeled data (Khanehzar et al. 2021; Chen et al. 2021a, b). ASM2TV (Chen et al. 2021a) designed a semi-supervised learning algorithm for fragmented time series that utilizes a large amount of unlabeled data to improve model performance. In weakly supervised learning, data labels are usually low quality. In the AVVP task, video-level labels are used for training, and precise labels are used at test time (Tian et al. 2020; Wu and Yang 2021; Yu et al. 2021a).

Unsupervised, weakly supervised, etc., will continue to be researched in the future to solve the problem of manual multi-view data annotations.

6.4 Unaligned multi-view sequence learning

As mentioned in Sect. 2.2.5, multi-view sequence asynchronous (sequence length is not equal or semantic misalignment) is common in real applications. Therefore, ignoring the asynchrony between these sequences will hinder subsequent tasks.

Multi-view alignment is performed by using pre-trained models. Aytar et al. (2017) provided a model for downstream tasks, using a large amount of synchronized data to learn three modes (visual, sound, and language) aligned and modally robust deep representations. In addition, the attention mechanism provides a feasible solution for sequence alignment and cross-view alignment (Tian et al. 2020; Le et al. 2018). Le et al. (2018) designed a memory-augmented network to model the interaction between two unaligned sequences. Some works implements asynchronous sequence alignment by using Transformer (Tsai et al. 2019; Delbrouck et al. 2020). Delbrouck et al. (2020) investigated a Transformer-based joint encoding method to jointly encode one or more modalities, which established global dependencies between input and output through an attention mechanism. The study of unaligned multi-view sequences is no longer limited by prior manual data alignment, and will be further applied to practical scenarios.

6.5 Trusted multi-view learning

MvSD is collected from different data sources. Various sensors or environmental factors may affect the quality of these views and bring noise. Analyzing these low-quality data, especially when unreliable views are presented, will seriously hinder multi-view tasks. In addition, for a specific task, the value of information expressed by multiple views is different, so the weight of each view is not fixed. Therefore, uncertainty estimation of MvSD is helpful to improve the robustness of multi-view.

Han et al. (2021d) proposed a unified trusted multi-view classification framework that applied a Dirichlet distribution to model the probability of each class and parameterized evidence from different views to estimate the uncertainty of each view. Finally, the Dempster Shafer theory was used to integrate the multi-view opinions. Geng et al. (2021) designed an unsupervised multi-view learning method that estimated views quality online through uncertainty modeling and integrated inherent information from multiple views to obtain a noise-free representation, thereby reducing the impact of quality imbalances of different views. Wang et al. (2019b) studied a negative log-likelihood error loss, which achieved single-value prediction and uncertainty quantification simultaneously. It predicted the mean and variance of the parameterized Gaussian distribution at each time step. Through uncertainty estimation, the model utilizes valuable information as much as possible and reduces the impact of low-quality views.

7 Conclusions

In this paper, we review the latest deep learning techniques in MvSD. We introduce four common data types that make up MvSD, including point data, sequence data, graph data, and raster data. We also enumerate the technical challenges of MvSD: temporal dynamic,

heterogeneity, cross-view dynamics, data missing and misalignment of asynchronous views. In addition, we summarize the representation methods of different data types in neural networks. Further, we reviewe the latest deep learning technology applied in MvSD. We also summarize some application areas of MvSD, and finally give several potential research directions in the future.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.61976247) and Sichuan Science and Technology Program (No.2021YFG0312).

References

Abavisani M, Naghizadeh A, Metaxas D, Patel V (2020) Deep subspace clustering with data augmentation. Adv Neural Inf Process Syst 33:10360–10370

Agyemang B, Wu W-P, Kpiebaareh MY, Lei Z, Nanor E, Chen L (2020) Multi-view self-attention for interpretable drug-target interaction prediction. J Biomed Inform 110:103547

Akhtar MS, Chauhan DS, Ghosal D, Poria S, Ekbal A, Bhattacharyya P (2019) Multi-task learning for multi-modal emotion recognition and sentiment analysis. In: NAACL-HLT (1)

Akman A, Coppock H, Gaskell A, Tzirakis P, Jones L, Schuller BW (2021) Evaluating the covid-19 identification resnet (cider) on the interspeech covid-19 from audio challenges. https://arXiv.org/2107.14549

Alam MM, Torgo L, Bifet A (2021) A survey on spatio-temporal data analytics systems. https://arXiv.org/2103.09883

Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In: International Conference on Machine Learning. PMLR, pp 1247–1255

Atluri G, Karpatne A, Kumar V (2018) Spatio-temporal data mining: a survey of problems and methods. ACM Comput Surv (CSUR) 51(4):1–41

Aytar Y, Vondrick C, Torralba A (2017) See, hear, and read: Deep aligned representations. https://arXiv.org/1706.00932

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. https://arXiv.org/1409.0473

Bai L, Yao L, Kanhere SS, Wang X, Liu W, Yang Z (2019) Spatio-temporal graph convolutional and recurrent networks for citywide passenger demand prediction. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp 2293–2296

Bai L, Yao L, Li C, Wang X, Wang C (2020) Adaptive graph convolutional recurrent network for traffic forecasting. Adv Neural Inf Process Syst 33:17804–17815

Bai J, Zhu J, Song Y, Zhao L, Hou Z, Du R, Li H (2021) A3t-gcn: attention temporal graph convolutional network for traffic forecasting. ISPRS Int J Geo-Inf 10(7):485

Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443

Barraza-Barraza D, Tercero-Gómez VG, Beruvides MG, Limón-Robles J (2017) An adaptive arx model to estimate the rul of aluminum plates based on its crack growth. Mech Syst Signal Process 82:519–536

Bie Y, Yang Y (2021) A multitask multiview neural network for end-to-end aspect-based sentiment analysis. Big Data Min Anal 4(3):195–207

Brousmiche M, Rouat J, Dupont S (2021) Multi-level attention fusion network for audio-visual event recognition. https://arXiv.org/2106.06736

Cai Y, Zeng M, Cai Z, Liu X, Zhang Z (2021) Graph regularized residual subspace clustering network for hyperspectral image clustering. Inf Sci 578:85–101

Chang C-C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):1–27

Chen C, Hu J, Meng Q, Zhang Y (2011) Short-time traffic flow prediction with arima-garch model. In: 2011 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp 607–612

Chen M, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency L-P (2017) Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp 163–171

Chen J, Xu X, Wu Y, Zheng H (2018) Gc-lstm: Graph convolution embedded lstm for dynamic link prediction. https://arXiv.org/1812.04206

Chen C, Li K, Teo SG, Zou X, Wang K, Wang J, Zeng Z (2019) Gated residual recurrent graph neural networks for traffic prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 485–492
Chen C, Li K, Teo SG, Zou X, Li K, Zeng Z (2020a) Citywide traffic flow prediction based on multiple gated spatio-temporal convolutional neural networks. ACM Trans Knowl Discov from Data (TKDD) 14(4):1–23
Chen W, Chen L, Xie Y, Cao W, Gao Y, Feng X (2020b) Multi-range attentive bicomponent graph convolutional network for traffic forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 3529–3536
Chen W, Wang W, Liu L, Lew MS (2020c) New ideas and trends in deep multimodal content understanding: a review. https://arXiv.org/2010.08189
Chen Z, Shi M, Zhang X, Ying H (2021a) Asm2tv: An adaptive semi-supervised multi-task multi-view learning framework. https://arXiv.org/2105.08643
Chen M, Du Y, Zhang Y, Qian S, Wang C (2021b) Semi-supervised learning with multi-head co-training. https://arXiv.org/2107.04795
Chen L, Xu J, Wu B, Qian Y, Du Z, Li Y, Zhang Y (2021c) Group-aware graph neural network for nationwide city air quality forecasting. https://arXiv.org/2108.12238
Cheng J, Dong L, Lapata M (2016) Long short-term memory-networks for machine reading. arXiv preprint https://arXiv.org/1601.06733
Cheng W, Shen Y, Zhu Y, Huang L (2018) A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In: Thirty-second AAAI Conference on Artificial Intelligence
Choe J, Im S, Rameau F, Kang M, Kweon IS (2021) Volumefusion: Deep depth fusion for 3d scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 16086–16095
Civitarese DS, Szwarcman D, Zadrozny B, Watson C (2021) Extreme precipitation seasonal forecast using a transformer neural network. https://arXiv.org/2107.06846
Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller BW (2021) End-2-end covid-19 detection from breath & cough audio. https://arXiv.org/2102.08359
Delbrouck J-B, Tits N, Brousmiche M, Dupont S (2020) A transformer-based joint-encoding for emotion recognition and sentiment analysis. https://arXiv.org/2006.15955
Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. https://arXiv.org/1810.04805
Du S, Li T, Yang Y, Horng S-J (2019) Deep air quality forecasting using hybrid deep learning framework. IEEE Trans Knowl Data Eng 33:2412
Du Y, Wang J, Feng W, Pan S, Qin T, Xu R, Wang C (2021) Adarnn: Adaptive learning and forecasting of time series. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp 402–411
Dumpala SH, Sheikh I, Chakraborty R, Kopparapu SK (2019) Audio-visual fusion for sentiment classification using cross-modal autoencoder. In: 32nd Conference on Neural Information Processing Systems (NIPS 2018), pp 1–4
Ertugrul AM, Lin Y-R, Taskaya-Temizel T (2019) Castnet: Community-attentive spatio-temporal networks for opioid overdose forecasting. https://arXiv.org/1905.04714
Fadadu S, Pandey S, Hegde D, Shi Y, Chou F-C, Djuric N, Vallespi-Gonzalez C (2022) Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2349–2357
Fang S, Zhang Q, Meng G, Xiang S, Pan C (2019) Gstnet: global spatial-temporal network for traffic flow prediction. In: IJCAI, pp 2286–2293
Fawagreh K, Gaber MM, Elyan E (2014) Random forests: from early developments to recent advancements. Syst Sci Control Eng Open Access J 2(1):602–609
Feng C-M, Yan Y, Chen G, Fu H, Xu Y, Shao L (2021) Accelerated multi-modal mr imaging with transformers. https://arXiv.org/2106.14248
Ferenstein E, Gasowski M (2004) Modelling stock returns with ar-garch processes. SORT Stat Oper Res Trans 28:55–68
Geng X, Li Y, Wang L, Zhang L, Yang Q, Ye J, Liu Y (2019) Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 3656–3663
Geng Y, Han Z, Zhang C, Hu Q (2021) Uncertainty-aware multi-view representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp 7545–7553
Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1440–1448

Gu Y, Yang K, Fu S, Chen S, Li X, Marsic I (2018) Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2018. NIH Public Access, p 2225

Guillot A, Thorey V (2021) Robustsleepnet: transfer learning for automated sleep staging at scale. IEEE Trans Neural Syst Rehabil Eng 29:1441–1451

Guo S, Lin Y, Li S, Chen Z, Wan H (2019a) Deep spatial-temporal 3d convolutional neural networks for traffic data forecasting. IEEE Trans Intell Transp Syst 20(10):3913–3926

Guo S, Lin Y, Feng N, Song C, Wan H (2019b) Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 922–929

Guo K, Hu Y, Qian Z, Sun Y, Gao J, Yin B (2020) Dynamic graph convolution network for traffic forecasting based on latent network of Laplace matrix estimation. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2020.3019497

Hackel T, Savinov N, Ladicky L, Wegner JD, Schindler K, Pollefeys M (2017) SEMANTIC3D.NET: a new large-scale point cloud classification benchmark. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-1-W1, pp 91–98

Han W, Chen H, Gelbukh A, Zadeh A, Morency L-P, Poria S (2021a) Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: Proceedings of the 2021 International Conference on Multimodal Interaction, pp 6–15

Han W, Chen H, Poria S (2021b) Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 9180–9192

Han Q, Lu D, Chen R (2021c) Fine-grained air quality inference via multi-channel attention model. In: IJCAI, pp 2512–2518

Han Z, Zhang C, Fu H, Zhou JT (2021d) Trusted multi-view classification. https://arXiv.org/2102.02051

Hasan MK, Lee S, Rahman W, Zadeh A, Mihalcea R, Morency L-P, Hoque E (2021) Humor knowledge enriched transformer for understanding multimodal humor. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp 12972–12980

Hazarika D, Zimmermann R, Poria S (2020) Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 1122–1131

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2961–2969

He K, Fan H, Wu Y, Xie S, Girshick R (2020a) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9729–9738

He Y, Wang C, Li N, Zeng Z (2020b) Attention and memory-augmented networks for dual-view sequential learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 125–134

Hotelling H (1992) Relations between two sets of variates. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics. Springer, Berlin, pp 162–190

Huang W, Song G, Hong H, Xie K (2014) Deep architecture for traffic flow prediction: deep belief networks with multitask learning. IEEE Trans Intell Transp Syst 15(5):2191–2201

Huang C, Zhang J, Zheng Y, Chawla NV (2018) Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp 1423–1432

Huang S, Kang Z, Xu Z (2020a) Auto-weighted multi-view clustering via deep matrix decomposition. Pattern Recogn 97:107015

Huang R, Huang C, Liu Y, Dai G, Kong W (2020b) Lsgcn: long short-term traffic prediction with graph convolutional networks. In: IJCAI, pp 2355–2361

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp 448–456

Islam T, Goldwasser D (2021) Twitter user representation using weakly supervised graph embedding. https://arXiv.org/2108.08988

Ismail AA, Hasan M, Ishtiaq F (2020) Improving multimodal accuracy through modality pre-training and attention. https://arXiv.org/2011.06102

Janjua PZ, Samad G, Khan N (2014) Climate change and wheat production in Pakistan: an autoregressive distributed lag approach. NJAS Wageningen J Life Sci 68:13–19

Ji P, Zhang T, Li H, Salzmann M, Reid I (2017) Deep subspace clustering networks. Adv Neural Inf Process Syst 30

Jia Z, Lin Y, Wang J, Zhou R, Ning X, He Y, Zhao Y (2020) Graphsleepnet: adaptive spatial-temporal graph convolutional networks for sleep stage classification. In: IJCAI, pp 1324–1330

Jia Z, Lin Y, Wang J, Ning X, He Y, Zhou R, Zhou Y, Li-wei HL (2021a) Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. IEEE Trans Neural Syst Rehabil Eng 29:1977–1986

Jia Z, Lin Y, Wang J, Wang X, Xie P, Zhang Y (2021b) Salientsleepnet: Multimodal salient wave detection network for sleep staging. https://arXiv.org/2105.13864

Kan M, Shan S, Chen X (2016) Multi-view deep network for cross-view classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4847–4855

Kemp B, Zwinderman AH, Tuk B, Kamphuisen HA, Oberye JJ (2000) Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. IEEE Trans Biomed Eng 47(9):1185–1194

Khan GA, Hu J, Li T, Diallo B, Zhao Y (2022a) Multi-view low rank sparse representation method for three-way clustering. Int J Mach Learn Cybern 13(1):233–253

Khan GA, Hu J, Li T, Diallo B, Wang H (2022b) Multi-view data clustering via non-negative matrix factorization with manifold regularization. Int J Mach Learn Cybern 13(3):677–689

Khanehzar S, Cohn T, Mikolajczak G, Turpin A, Frermann L (2021) Framing unpacked: A semi-supervised interpretable multi-view model of media frames. https://arXiv.org/2104.11030

Kim D, Tsai Y-H, Zhuang B, Yu X, Sclaroff S, Saenko K, Chandraker M (2021) Learning cross-modal contrastive features for video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 13618–13627

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. https://arXiv.org/1609.02907

Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: International Conference on Machine Learning. PMLR, pp 595–603

Laddha A, Gautam S, Palombo S, Pandey S, Vallespi-Gonzalez C (2021) Mvfusenet: Improving end-to-end object detection and motion forecasting through multi-view fusion of lidar data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2865–2874

Le H, Tran T, Venkatesh S (2018) Dual memory neural computer for asynchronous two-view sequential learning. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 1637–1645

Liang PP, Liu Z, Zadeh A, Morency L-P (2018a) Multimodal language analysis with recurrent multistage fusion. https://arxiv.org/1808.03920

Liang PP, Zadeh A, Morency LP (2018b) Multimodal local-global ranking fusion for emotion recognition. In: the 2018

Liang Y, Ouyang K, Jing L, Ruan S, Liu Y, Zhang J, Rosenblum DS, Zheng Y (2019) Urbanfm: Inferring fine-grained urban flows. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 3132–3142

Liao B, Zhang J, Wu C, McIlwraith D, Chen T, Yang S, Guo Y, Wu F (2018) Deep sequence learning with auxiliary information for traffic prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 537–546

Li Y, Moura JM (2019) Forecaster: a graph transformer for forecasting spatial and time-dependent data. https://arXiv.org/1909.04019

Li Y, Yu R, Shahabi C, Liu Y (2017) Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. https://arXiv.org/1707.01926

Li Z, Wang Q, Tao Z, Gao Q, Yang Z, et al (2019) Deep adversarial multi-view clustering network. In: IJCAI, pp 2952–2958

Li Z, Wang H, Li J (2020) Auto-mvcnn: neural architecture search for multi-view 3d shape recognition. https://arXiv.org/2012.05493

Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2117–2125

Lin Y, Chiang Y-Y, Franklin M, Eckel SP, Ambite JL (2020) Building autocorrelation-aware representations for fine-scale spatiotemporal prediction. In: 2020 IEEE International Conference on Data Mining (ICDM). IEEE, pp 352–361

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016a) Ssd: single shot multibox detector. In: European Conference on Computer Vision. Springer, pp 21–37

Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016b) Urban water quality prediction based on multi-task multi-view learning
Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency L-P (2018) Efficient low-rank multimodal fusion with modality-specific factors. https://arXiv.org/1806.00064
Liu S, Fan H, Qian S, Chen Y, Ding W, Wang Z (2021) Hit: Hierarchical transformer with momentum contrast for video-text retrieval. https://arXiv.org/2103.15049
Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440
Lu GY, Wong DW (2008) An adaptive inverse-distance weighting spatial interpolation technique. Comput Geosci 34(9):1044–1055
Lu Y-J, Li C-T (2020) Agstn: learning attention-adjusted graph spatio-temporal networks for short-term urban sensor value forecasting. In: 2020 IEEE International Conference on Data Mining (ICDM). IEEE, pp 1148–1153
Lv J, Sun Q, Li Q, Moreira-Matias L (2019) Multi-scale and multi-scope convolutional neural networks for destination prediction of trajectories. IEEE Trans Intell Transp Syst 21(8):3184–3195
Ma T, Xiao C, Zhou J, Wang F (2018) Drug similarity integration through attentive multi-view graph auto-encoders. https://arXiv.org/1804.10850
Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., & Peng, X. (2021). SMIL: Multimodal Learning with Severely Missing Modality. Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), 2302–2310
Mai S, Xing S, Hu H (2019) Locally confined modality fusion network with a global perspective for multimodal human affective computing. IEEE Trans Multimedia 22(1):122–137
Mai S, Hu H, Xing S (2020) Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 164–172
Mai S, Zeng Y, Zheng S, Hu H (2021) Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. https://arXiv.org/2109.01797
Mao L, Sun S (2020) Multiview variational sparse gaussian processes. IEEE Trans Neural Netw Learn Syst 32(7):2875–2885
Mazimpaka JD, Timpf S (2016) Trajectory data mining: a review of methods and applications. J Spatial Inf Sci 2016(13):61–99
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. https://arXiv.org/1301.3781
Nessiem MA, Mohamed MM, Coppock H, Gaskell A, Schuller BW (2021) Detecting covid-19 from breathing and coughing sounds using deep neural networks. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, pp 183–188
Okawa M, Iwata T, Kurashima T, Tanaka Y, Toda H, Ueda N (2019) Deep mixture point processes: spatio-temporal event prediction with rich contextual information. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 373–383
Olesen AN, Jennum P, Mignot E, Sorensen HB (2021) Msed: a multi-modal sleep event detection model for clinical sleep analysis. https://arXiv.org/2101.02530
Ordóñez C, Lasheras FS, Roca-Pardiñas J, de Cos Juez FJ (2019) A hybrid arima-svm model for the study of the remaining useful life of aircraft engines. J Comput Appl Math 346:184–191
Ouyang X, Yang Y, Zhang Y, Zhou W (2021) Spatial-temporal dynamic graph convolution neural network for air quality prediction. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1–8
Pan Z, Liang Y, Wang W, Yu Y, Zheng Y, Zhang J (2019) Urban traffic prediction from spatio-temporal data using deep meta learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 1720–1730
Patel ZB, Purohit P, Patel HM, Sahni S, Batra N (2022) Accurate and scalable gaussian processes for fine-grained air quality inference. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp 12080–12088
Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
Pérez-Rúa J-M, Vielzeuf V, Pateux S, Baccouche M, Jurie F (2019) Mfas: multimodal fusion architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6966–6975
Pham HT, Yang B-S (2010) Estimation and forecasting of machine health condition using arma/garch model. Mech Syst Signal Process 24(2):546–558

Pham H, Manzini T, Liang PP, Poczos B (2018) Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. https://arXiv.org/1807.03915

Pham H, Liang PP, Manzini T, Morency L-P, Póczos B (2019) Found in translation: Learning robust joint representations by cyclic translations between modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 6892–6899

Phan H, Andreotti F, Cooray N, Chén OY, De Vos M (2019) Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng 27(3):400–410

Phan H, Chén OY, Tran MC, Koch P, Mertins A, De Vos M (2021) Xsleepnet: multi-view sequential model for automatic sleep staging. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3070057

Phan H, Mikkelsen KB, Chen O, Koch P, Mertins A, De Vos M (2022) Sleeptransformer: automatic sleep staging with interpretability and uncertainty quantification. IEEE Trans Biomed Eng 69:2456

Piriyajitakonkij M, Warin P, Lakhan P, Leelaarporn P, Kumchaiseemak N, Suwajanakorn S, Pianpanit T, Niparnan N, Mukhopadhyay SC, Wilaiprasitporn T (2020) Sleepposenet: multi-view learning for sleep postural transition recognition using uwb. IEEE J Biomed Health Inform 25(4):1305–1314

Poria S, Cambria E, Hazarika D, Mazumder N, Zadeh A, Morency L-P (2017) Multi-level multiple attentions for contextual multimodal sentiment analysis. In: 2017 IEEE International Conference on Data Mining (ICDM). IEEE, pp 1033–1038

Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M (2020) U2-net: going deeper with nested u-structure for salient object detection. Pattern Recogn 106:107404

Rahate A, Walambe R, Ramanna S, Kotecha K (2021) Multimodal co-learning: challenges, applications with datasets, recent advances and future directions. https://arXiv.org/2107.13782

Rahman W, Hasan MK, Lee S, Zadeh A, Mao C, Morency L-P, Hoque E (2020) Integrating multimodal information in large pretrained transformers. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2020. NIH Public Access, p 2359

Ramachandram D, Taylor GW (2017) Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 34(6):96–108

Rayhan Y, Hashem T (2020) Aist: An interpretable attention-based deep learning model for crime prediction. https://arxiv.org/arXiv:2012.08713

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 779–788

Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99

Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp 234–241

Salama U, Chen X, Yao L, Paik H-Y, Wang X (2021) Deep multi-view spatio-temporal network for urban crime prediction. In: Australasian Database Conference. Springer, pp 50–61

Sasaki Y, Harada K, Yamasaki S, Onizuka M (2021) Airex: Neural network-based approach for air quality inference in unmonitored cities. https://arXiv.org/2108.07120

Shi X, Qi H, Shen Y, Wu G, Yin B (2020) A spatial-temporal attention approach for traffic prediction. IEEE Trans Intell Transp Syst 22:4909

Song C, Lin Y, Guo S, Wan H (2020) Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 914–921

Stec A, Klabjan D (2018) Forecasting crime with deep learning. https://arXiv.org/1806.01486

Summaira J, Li X, Shoib AM, Li S, Abdul J (2021) Recent advances and trends in multimodal deep learning: a review. https://arXiv.org/2105.11087

Sun S, Zong D (2020) Lcbm: a multi-view probabilistic model for multi-label classification. IEEE Trans Pattern Anal Mach Intell 43(8):2682–2696

Sun S, Zhao J, Gao Q (2015) Modeling and recognizing human trajectories with beta process hidden Markov models. Pattern Recogn 48(8):2407–2417

Sun S, Wu H, Xiang L (2020a) City-wide traffic flow forecasting using a deep convolutional neural network. Sensors 20(2):421

Sun J, Zhang J, Li Q, Yi X, Liang Y, Zheng Y (2020b) Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2020.3008774

Sun S, Dong Z, Zhao J (2020c) Conditional random fields for multiview sequential data modeling. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2020.3041591

Sun S, Dong W, Liu Q (2020d) Multi-view representation learning with deep gaussian processes. IEEE Trans Pattern Anal Mach Intell 43(12):4453–4468
Sun Z, Sarma P, Sethares W, Liang Y (2020e) Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 8992–8999
Supratak A, Guo Y (2020) Tinysleepnet: an efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, pp 641–644
Supratak A, Dong H, Wu C, Guo Y (2017) Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. IEEE Trans Neural Syst Rehabil Eng 25(11):1998–2008
Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp 3104–3112
Taylor SJ, Letham B (2018) Forecasting at scale. Am Stat 72(1):37–45
Tian Y, Li D, Xu C (2020) Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, pp 436–454
Torres C, Fragoso V, Hammond SD, Fried JC, Manjunath B (2016) Eye-cu: Sleep pose classification for healthcare using multimodal multiview data. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 1–9
Tran L, Liu X, Zhou J, Jin R (2017) Missing modalities imputation via cascaded residual autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1405–1414
Tsai Y-HH, Liang PP, Zadeh A, Morency L-P, Salakhutdinov R (2018) Learning factorized multimodal representations. https://arXiv.org/1806.06176
Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R (2019) Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2019, p 6558. NIH Public Access
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, pp 5998–6008
Veiga T, Ljunggren E, Bach K, Akselsen S (2021) Blind calibration of air quality wireless sensor networks using deep neural networks. In: 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS). IEEE, pp 1–6
Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. https://arXiv.org/1710.10903
Verma S, Wang J, Ge Z, Shen R, Jin F, Wang Y, Chen F, Liu W (2020) Deep-hoseq: deep higher order sequence fusion for multimodal sentiment analysis. In: 2020 IEEE International Conference on Data Mining (ICDM). IEEE, pp 561–570
Vomfell L, Härdle WK, Lessmann S (2018) Improving crime count forecasts using twitter and taxi data. Decis Support Syst 113:73–85
Wang Y (2021) Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. ACM Trans Multimedia Comput Commun Appl (TOMM) 17(1s):1–25
Wang W, Arora R, Livescu K, Bilmes J (2015) On deep multi-view representation learning. In: International Conference on Machine Learning. PMLR, pp 1083–1092
Wang D, Cao W, Li J, Ye J (2017) Deepsd: Supply-demand prediction for online car-hailing services using deep neural networks. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp 243–254. IEEE
Wang B, Yin P, Bertozzi AL, Brantingham PJ, Osher SJ, Xin J (2019a) Deep learning for real-time crime forecasting and its ternarization. Chin Ann Math Ser B 40(6):949–966
Wang B, Lu J, Yan Z, Luo H, Li T, Zheng Y, Zhang G (2019b) Deep uncertainty quantification: A machine learning approach for weather forecasting. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2087–2095
Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency L-P (2019c) Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 7216–7223
Wang J, Wang W, Wang L, Wang Z, Feng DD, Tan T (2020a) Learning visual relationship and context-aware attention for image captioning. Pattern Recogn 98:107075
Wang Q, Cheng J, Gao Q, Zhao G, Jiao L (2020b) Deep multi-view subspace clustering with unified and discriminative learning. IEEE Trans Multimedia 23:3483–3493
Wang Q, Lian H, Sun G, Gao Q, Jiao L (2020c) Icmsc: incomplete cross-modal subspace clustering. IEEE Trans Image Process 30:305–317

Wang S, Cao J, Yu P (2020d) Deep learning for spatio-temporal data mining: a survey. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2020.3025580
Wang X, Ma Y, Wang Y, Jin W, Yu J (2020e) Traffic flow prediction via spatial temporal graph neural network. In: WWW '20: The Web Conference 2020
Wang S, Miao H, Chen H, Huang Z (2020f) Multi-task adversarial spatial-temporal networks for crowd flow prediction. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp 1555–1564
Wang Z, Wan Z, Wan X (2020g) Transmodality: an end2end fusion method with transformer for multimodal sentiment analysis. In: Proceedings of The Web Conference 2020, pp 2514–2520
Wang C, Lin Z, Yang X, Sun J, Yue M, Shahabi C (2021) Hagen: Homophily-aware graph convolutional recurrent network for crime forecasting. https://arXiv.org/2109.12846
Wang J, Yang Y, Liu K, Xie P, Liu X (2022) Instance-guided multi-modal fake news detection with dynamic intra- and inter-modality fusion. In: PAKDD, pp 510–521
Wu Y, Yang Y (2021) Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1326–1335
Wu X, Huang C, Zhang C, Chawla NV (2020) Hierarchically structured transformer networks for fine-grained spatial event forecasting. In: Proceedings of The Web Conference 2020, pp 2320–2330
Wu J, Jiang Z, Wen S, Men A, Wang H (2021) Rethinking the constraints of multimodal fusion: case study in weakly-supervised audio-visual video parsing. https://arXiv.org/2105.14430
Xia L, Huang C, Xu Y, Dai P, Bo L, Zhang X, Chen T (2021) Spatial-temporal sequential hypergraph network for crime prediction with dynamic multiplex relation learning. In: IJCAI, pp 1631–1637
Xia W, Wang S, Yang M, Gao Q, Han J, Gao X (2022) Multi-view graph embedding clustering network: joint self-supervision and block diagonal representation. Neural Netw 145:1–9
Xingjian S, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, pp 802–810
Xu H, Zhang H, Han K, Wang Y, Peng Y, Li X (2019) Learning alignment for multimodal emotion recognition from speech. https://arXiv.org/1909.05645
Xu Z, So DR, Dai AM (2021) Mufasa: Multimodal fusion architecture search for electronic health records. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp 10532–10540
Yang H-F, Dillon TS, Chen Y-PP (2016) Optimized structure of the traffic flow forecasting model with a deep learning approach. IEEE Trans Neural Netw Learn Syst 28(10):2371–2381
Yang B, Luo W, Urtasun R (2018) Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7652–7660
Yang J, Wang Y, Yi R, Zhu Y, Rehman A, Zadeh A, Poria S, Morency L-P (2020) Mtgat: multimodal temporal graph attention networks for unaligned human multimodal language sequences. https://arXiv.org/2010.11985
Yao H, Wu F, Ke J, Tang X, Jia Y, Lu S, Gong P, Ye J, Li Z (2018) Deep multi-view spatial-temporal network for taxi demand prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
Yao H, Liu Y, Wei Y, Tang X, Li Z (2019a) Learning from multiple cities: a meta-learning approach for spatial-temporal prediction. In: The World Wide Web Conference, pp 2181–2191
Yao H, Tang X, Wei H, Zheng G, Li Z (2019b) Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 5668–5675
Yi X, Zhang J, Wang Z, Li T, Zheng Y (2018) Deep distributed fusion network for air quality prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 965–973
Yin J, Sun S (2019) Multiview uncorrelated locality preserving projection. IEEE Trans Neural Netw Learn Syst 31(9):3442–3455
Yin X, Wu G, Wei J, Shen Y, Qi H, Yin B (2021a) Multi-stage attention spatial-temporal graph networks for traffic prediction. Neurocomputing 428:42–53
Yin Y, Huang S, Zhang X, Dou D (2021b) Bm-nas: Bilevel multimodal neural architecture search. https://arXiv.org/2104.09379
Yu B, Yin H, Zhu Z (2017) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. https://arXiv.org/1709.04875
Yu Z, Cui Y, Yu J, Wang M, Tao D, Tian Q (2020a) Deep multimodal neural architecture search. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 3743–3752

Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J, Zou J, Yang K (2020b) Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 3718–3727
Yu J, Cheng Y, Zhao R-W, Feng R, Zhang Y (2021a) Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. https://arXiv.org/2111.12374
Yu W, Xu H, Yuan Z, Wu J (2021b) Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp 10790–10797
Yuan Z, Zhou X, Yang T (2018) Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 984–992
Yuan Y, Jia K, Ma F, Xun G, Wang Y, Su L, Zhang A (2019) A hybrid self-attention deep learning framework for multivariate sleep stage classification. BMC Bioinform 20(16):1–10
Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. https://arXiv.org/1606.06259
Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. https://arXiv.org/1707.07250
Zadeh AB, Liang PP, Poria S, Cambria E, Morency L-P (2018a) Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp 2236–2246
Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P (2018b) Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency L-P (2018c) Multi-attention recurrent network for human communication comprehension. In: Thirty-Second AAAI Conference on Artificial Intelligence
Zadeh A, Mao C, Shi K, Zhang Y, Liang PP, Poria S, Morency L-P (2019) Factorized multimodal transformer for multimodal sequential learning. rint https://arXiv.org/1911.09826
Zhang J, Zheng Y, Qi D, Li R, Yi X (2016) Dnn-based prediction model for spatio-temporal data. In: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp 1–4
Zhang J, Zheng Y, Qi D (2017) Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Thirty-first AAAI Conference on Artificial Intelligence
Zhang G-Q, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S (2018) The national sleep research resource: towards a sleep data commons. J Am Med Inform Assoc 25(10):1351–1358
Zhang J, Zheng Y, Sun J, Qi D (2019) Flow prediction in spatio-temporal networks based on multitask deep learning. IEEE Trans Knowl Data Eng 32(3):468–478
Zhang C, Cui Y, Han Z, Zhou JT, Fu H, Hu Q (2020) Deep partial multi-view learning. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.3037734
Zhang Y, Yang Y, Zhou W, Wang H, Ouyang X (2021a) Multi-city traffic flow forecasting via multi-task learning. Appl Intell 51:6895
Zhang M, Li T, Li Y, Hui P (2021b) Multi-view joint graph representation learning for urban region embedding. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp 4431–4437
Zhao J, Sun S (2016a) High-order gaussian process dynamical models for traffic flow prediction. IEEE Trans Intell Transp Syst 17(7):2014–2019
Zhao J, Sun S (2016b) Variational dependent multi-output gaussian process dynamical systems. J Mach Learn Res 17(1):4134–4169
Zhao J, Xie X, Xu X, Sun S (2017a) Multi-view learning overview: recent progress and new challenges. Inf Fusion 38:43–54
Zhao H, Ding Z, Fu Y (2017b) Multi-view clustering via deep matrix factorization. In: Thirty-first AAAI Conference on Artificial Intelligence
Zhao H, Shi J, Qi X, Wang X, Jia J (2017c) Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2881–2890
Zhao L, Song Y, Zhang C, Liu Y, Wang P, Lin T, Deng M, Li H (2019) T-gcn: a temporal graph convolutional network for traffic prediction. IEEE Trans Intell Transp Syst 21(9):3848–3858
Zheng C, Fan X, Wang C, Qi J (2020) Gman: a graph multi-attention network for traffic prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 1234–1241

Zheng L, Cheng Y, Yang H, Cao N, He J (2021) Deep co-attention network for multi-view subspace learning. In: Proceedings of the Web Conference 2021, pp 1528–1539
Zhong H, Yin C, Wu X, Luo J, He J (2020) Airrl: A reinforcement learning approach to urban air quality inference. https://arXiv.org/2003.12205
Zhou Y, Tuzel O (2018) Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4490–4499
Zhou W, Yang Y, Zhang Y, Wang D, Zhang X (2020) Deep flexible structured spatial-temporal model for taxi capacity prediction. Knowl-Based Syst 205:106286