Dhanesh Ramachandram and
Graham W. Taylor

Deep Multimodal Learning

A survey on recent advances and trends

©ISTOCKPHOTO.COM/ZAPP2PHOTO

The success of deep learning has been a catalyst to solving increasingly complex machine-learning problems, which often involve multiple data modalities. We review recent advances in deep multimodal learning and highlight the state-of the art, as well as gaps and challenges in this active research field. We first classify deep multimodal learning architectures and then discuss methods to fuse learned multimodal representations in deep-learning architectures. We highlight two areas of research—regularization strategies and methods that learn or optimize multimodal fusion structures—as exciting areas for future work.

Introduction

Neural networks have made an impressive resurgence in recent years, after long-standing concerns about the ability to train deep models were successfully abated by a pioneering group of researchers who leveraged advances in algorithms, data, and computation [1]. This active research area now interests researchers in academia, but also industry, and it has resulted in state-of-the-art performance for many practical problems, especially in areas involving high-dimensional unstructured data such as in computer vision, speech, and natural language processing.

With the undeniable success of deep learning in the visual domain, the natural progression of deep-learning research points to problems involving larger and more complex multimodal data. Such multimodal data sets consist of data from different sensors observing a common phenomena, and the goal is to use the data in a complementary manner toward learning a complex task. One of the main advantages of deep learning is that a hierarchical representation can be automatically learned for each modality, instead of manually designing or handcrafting modality-specific features that are then fed to a machine-learning algorithm.

The goal of this article is to provide a comprehensive survey of the state of the art in deep multimodal learning and suggest future research directions by highlighting advances, gaps, and challenges in this active field. We believe this review is timely given the increasing number of deep-learning techniques

Digital Object Identifier 10.1109/MSP.2017.2738401
Date of publication: 13 November 2017

applied to multimodal data published in leading conferences and journals, as shown in Figure 1.

The crux of this article centers around two important areas of focus in deep multimodal learning research: 1) methods that use regularization techniques to improve cross-modality learning (see the section "Multimodal Regularization") and 2) methods that attempt to find optimal deep multimodal architectures through search, optimization, or some learning procedure (see the section "Fusion Structure Learning and Optimization").

Background

For the purposes of our review, we adopt the definition provided by Lahat et al. [2], where we consider phenomena or systems that are observed using multiple sensors and each sensor output can be termed a *modality* associated with a single data set. The underlying motivation to use multimodal data is that complementary information could be extracted from each of the modalities considered for a given learning task, yielding a richer representation that could be used to produce much improved performance compared to using only a single modality. There are many practical tasks that benefit from the use of multimodal data. In medical image analysis, for example, the use of multiple imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound imaging provides complementary information that is routinely used by medical experts in diagnosis and treatment. Applications involving human–computer interaction use depth and vision cues extensively for applications like immersive gaming and autonomous driving. Similar improvements in performance can be seen in biometric applications. In remote sensing applications, data from different sensors [intensity images, synthetic aperture radar, and light detection and ranging (LIDAR)] are often fused.

Techniques for multimodal data fusion, which cover different application domains, have long been investigated by the research community [3], [4]. Traditionally, combining the signals of multiple sensors has been investigated from a data fusion perspective. This is called *early fusion* or *data-level fusion* and focuses on how best to combine data from multiple sources, either by removing correlations between modalities or representing the fused data in a lower-dimensional common subspace. Techniques that accomplish one or both of these objectives include principal component analysis (PCA), independent components analysis, and canonical correlation analysis. The fused data are then presented to a machine-learning algorithm. When ensemble classifiers became popular in the early 2000s [5], researchers began applying multimodal fusion techniques that fell into the category known as *late fusion* or *decision-level fusion*. In general, these late-fusion strategies were much simpler to implement than early fusion, particularly when the different modalities varied significantly in terms of data dimensionality and sampling rates, and often resulted in improved performance.

As shown in the section "Intermediate Fusion," popular deep neural network (DNN) architectures allow yet a third form of multimodal fusion, i.e., intermediate fusion of learned representations, offering a truly flexible approach to multimodal fusion for numerous practical problems. As deep-learning architectures learn a hierarchical representation of the underlying data across its hidden layers, learned representations between different modalities can be fused at various levels of abstraction.

Deep-learning-based multimodal learning offers several advantages over conventional machine-learning methods, which are highlighted in Table 1. For many practical problems, deep-learning models often offer much improved performance for problems involving multimodal data. However, this entails several architectural design choices that we discuss next.

The first of these design choices relates to when to fuse different modalities. From a traditional data fusion standpoint, the practitioner could fuse the various input modalities at the data level and proceed to train a single machine-learning model, but, as we discuss in the section "Early Fusion," this option can be rather challenging. Alternatively, a late-fusion option can also be considered, and we review several works in this category in the section "Late Fusion." An important feature of deep learning is its ability to learn hierarchical representations from raw data. This feature can be exploited in multimodal learning to have a fine-grained control over how learned representations are fused. Therefore, a common practice in multimodal deep learning is to construct a shared representation or fusion layer that can merge incoming representations of modalities, thereby forcing the network to learn a joint representation of its inputs. The simplest fusion layer is a layer of hidden units, each of which receives input from all modalities. The flexibility of learning cross-modality shared representations at different levels of abstraction could be exploited to achieve better multimodal fusion results; however,

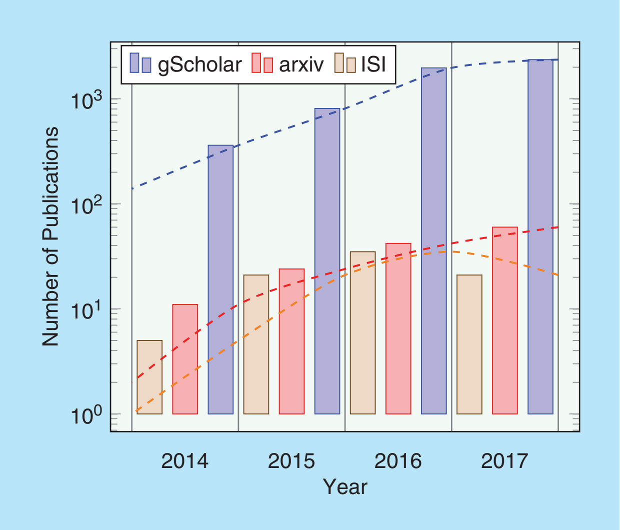

FIGURE 1. The increasing research interest in deep multimodal learning. Data were generated by analyzing the search results from leading search engines for technical publications: Google Scholar, the arXiv, and Thomson-Reuters' ISI. We used the search terms *multimodal*, *fusion,* and *deep learning* and applied search filters to include references from engineering, computer science, and mathematics fields, excluding results related to social sciences, neurobiology, and business. Note that, due to differences in scale in terms of returned results, we use a semilog scale.

Table 1. A comparison between deep multimodal learning and conventional approaches.

Deep Multimodal Learning	Conventional Multimodal Learning
Both modality-wise representations (features) and shared (fused) representations are learned from data.	Features are manually designed and require prior knowledge about the underlying problem and data.
Requires little or no preprocessing of input data (end-to-end training).	Some techniques, like early fusion, may be sensitive to data preprocessing.
Implicit dimensionality reduction within architecture.	Feature selection and dimensionality reduction are often explicitly performed.
Supports early, late, or intermediate fusion.	Typically performs early or late fusion.
Easily scalable in terms of data size and number of modalities for all fusion methods.	Early fusion (data-level fusion) can be challenging and not scalable: late-fusion rules may need to be defined.
Fusion architecture can be learned during training.	Rigid fusion architecture usually handcrafted.
Deeper, complex networks typically require large amounts of training data (if trained from scratch).	May not require as much training data.
Numerous hyperparameter tunings vital for state-of-art performance.	May not have as many hyperparameters as deep-learning architectures.
Compute intensive, requires powerful graphics processing units (GPUs) for reasonable training time.	GPUs may offer speed-up, but not critical.

the question remains: at which depth of representation would the fusion be optimal?

The second architectural design choice for deep multimodal learning concerns which modalities to fuse. The underlying assumption in multimodal fusion is that different modalities provide complementary information toward solving the task at hand. However, it could be the case that the inclusion of all available modalities ends up being detrimental to the performance of the machine-learning algorithm—and, as such, some form of feature selection may be required. In the section “Fusion Structure Learning and Optimization,” we discuss a number of techniques that, during training, can automatically learn the optimal ordering and depth of fusion.

The third design choice involves dealing with missing data or modalities. Deep multimodal learning models should be robust enough to compensate for missing data or modalities during inference. Generative models are typically used in such instances.

Most deep multimodal learning approaches also involve representation learning from raw data. It is often the case that a deep multimodal architecture utilizes several standard modules or “building blocks” that are optimized for a specific kind of data. The choice of which deep-learning module is best for extracting pertinent information for a given modality is also an important architectural design choice. For example, when two-dimensional (2-D) pixel-based data are considered, convolutional architectures are often preferred. Three-dimensional (3-D) convolutional networks can be used for volumetric data, like CT, MRI, or even video. When temporal data are used, variants of recurrent neural networks (RNNs) such as long short-term memory (LSTM) or gated recurrent units can be incorporated.

The choice of modality-wise deep-learning architecture is mainly dependent upon the dimensionality of the input or whether temporal trends need to be learned. Beyond these common architectural choices, it is up to the reader to decide, given application-specific requirements that may involve properties of the data set or even the hardware used for training or deployment.

Applications

This section aims to provide an overview of the various application domains where deep multimodal learning has garnered much interest. Although multimodal learning and fusion is a widely researched topic, deep multimodal learning only started to gain attention following the works of Ngiam et al. [6] and Srivastava and Salakhutdinov [7]. These early works on deep multimodal fusion involved only two modalities: images and text. Ngiam et al. [6] investigated several approaches for multimodal fusion that include simple concatenation of inputs and shared representation learning, as well as cross-modality learning (where data from all modalities are present during training, but only a single modality is available during test). At around the same time, Srivastava and Salakhutdinov [7] also demonstrated the utility of fusing higher-level representations of disparate modalities involving images and text in a deep-learning framework. A notable finding was that constructing a multimodal fusion layer by way of simple concatenation of incoming connections resulted in relatively poorer results—revealing that hidden units have strong connections to variables from individual modalities but few units that connect across modalities. They also found that capturing cross-modality correlations required at least one nonlinear stage to be successful since higher-level representations of individual modalities will be relatively “modality free” and therefore more amenable to fusion. These early explorations became the basis of a number of proceeding works in deep multimodal learning that investigated various regularization strategies (see the section “Multimodal Regularization”) to enforce constraints for learning intermodality relationships.

Human activity recognition

An important area of research that heavily utilizes multimodal data is human activity recognition. Under this large umbrella of research, there are numerous subfields of research that relate to some aspect of human understanding. Given that humans

Table 2. Multimodal learning data sets and public multimodal machine-learning challenges.

Data Set	Modalities	Problem	Reference	Year
UTD-MHAD	Depth and inertial sensor data	Human action recognition	Chen et al. [93]	2015
ChaLearn looking at people	RGB-D, audio, skeletal pose	Human activity recognition	Escalera et al. [94]	2014
Berkeley MHAD	Multiviewpoint RGB-D and skeletal pose data	Human activity recognition	Ofli et al. [95]	2013
MHRI data set	Chest, top RGB-D, face, video, and audio	Human–robot interaction	Pablo et al. [96]	2016
H-MOG	Nine smartphone sensors and interaction data	Continuous authentication in smartphones	Sitová et al. [12]	2016
RECOLA	Audio, visual, and physiological	Emotion recognition	Ringeval et al. [97]	2013
MHEALTH	Accelerometer, electrocardiogram, magnetometer, and gyroscopes	Health monitoring	Banos et al. [98]	2015
Pinterest Multimodal	Images and text (40M)	Multimodal word embeddings	Mao et al. [99]	2016
MM-IMDb	Video, images, and text metadata	Movie genre prediction	Arevalo et al. [100]	2017
FCVID	Video and audio	Action recognition	Jiang et al. [101]	2017
KITTI	Stereo gray- and color video, 3-D-LIDAR, inertial and GPS navigation data	Autonomous driving	Geiger et al. [91]	2017
KinectFaceDB	RGB-D and facial landmarks	Face recognition	Min et al. [102]	2014
Oxford RobotCar	Six cameras, LIDAR, GPS, and inertial navigation data	Autonomous driving	Maddern et al. [92]	2016
Multimodal BRATS	T2-, FLAIR-, post-Gadolinium T1-MRI, perfusion, and diffusion MRI and MRSI	Brain tumor segmentation	Menze et al. [103]	2015

RGB-D: RGB-depth

exhibit highly complex behavior in social settings, it is only natural that multimodal data are required for machine-learning algorithms to classify, or “understand,” their human behavior. Not surprisingly, we found that many works in deep multimodal fusion reported in recent years have focused on multimedia data that typically involve modalities such as audio, video, depth, and skeletal motion information. Multimodal deep-learning methods have been applied to various problems involving human activity such as action recognition (an activity can be composed of two or more sequences of shorter actions), gesture recognition [8], gaze-direction estimation [9], face recognition [10], and emotion recognition [11]. The ubiquity of mobile smartphones, which have no fewer than ten sensors, has given rise to novel applications that involve multimodal data such as continuous biometric authentication [12] and activity recognition [13]. Related subfields of research include human pose estimation [14] and semantic scene understanding [15].

We expect that the deep-learning research community will continue to focus on these problems in the foreseeable future. This is evidenced not only by the number of multimodal deep-learning papers being published but also the increasing number of data sets and public challenges made available online (see Table 2).

Medical applications

Deep learning in medical applications has become an important application domain that has attracted substantial interest. Medical imaging, for example, consists of a multitude of multimodal data in the form of different medical imaging modalities such as MRI, CT, positron emission tomography (PET), functional MRI (fMRI), X-ray, and ultrasound. Although there have been notable improvements in new medical imaging technologies, the interpretation of these modalities for diagnosis still requires highly trained human experts. Conventional computer vision approaches required manually designed morphological feature representations. However, transforming the tacit knowledge of human experts into a computational form is not trivial. In the medical imaging field, manually designing suitable image features is extremely challenging, as it not only involves the interpretation of subtle visual markers and abnormalities, often needing years of medical training, but also the need to fuse complementary as well as possibly conflicting information from multiple imaging modalities. Therefore, the ability to learn these multimodal features through examples, as seen in the success of deep learning applied to computer vision, has prompted researchers to investigate their applicability in the medical domain. It is therefore not surprising that an increasing amount of medical image analysis research in recent years [16], both unimodal as well as multimodal, involves deep-learning-based methods.

Multimodal deep learning has been used in problems involving tissue and organ segmentation [17], multimodal medical image retrieval [18], multimodal medical image registration [19], and computer-aided diagnosis [20]. A recent review article by Mamoshina et al. [21] demonstrates the rising popularity of DNNs, including models that implement multimodal fusion in biomedical applications involving genomic, proteomic, and drug data.

Two major challenges in applying deep-learning-based approaches for medical applications are 1) the difficulty in obtaining sufficiently labeled data and 2) the problem of class imbalance (where the number of negative examples far outnumber the case of positive examples). To overcome the first challenge, early approaches resorted to patch-based training [22]. Recently, techniques that utilize transfer learning have been surprisingly successful [23]. This involves reusing a portion of the data-agnostic representations learned by very deep networks on a separate, large data set, for example, ImageNet, and then fine-tuning or retraining only the upper layers of the network using a much smaller medical data set. Another common technique is to perform training data augmentation, for example, applying different affine transformations or randomly perturbing the brightness and contrast of images to increase the amount of training data available. To address the data imbalance problem, it is common to apply some form of weighting to the loss function such that mistakes made on the majority classes are less penalized than mistakes the network makes on the minority class. These challenges, although very common to problems in the medical domain, may also occur in other domains—and the solutions, as such, are equally applicable. However, despite the success of deep learning in medical applications, the medical community is still rather apprehensive about deploying them in the real world, as deep learning is often seen as opaque. This view is likely to gradually change given the increasing efforts to design interpretable DNNs [24], [25].

Autonomous systems

Following the success of deep learning, there has been a surge of interest in autonomous navigation (also known as *autonomous driving*) applications, which typically involve multimodal data acquired from sensors mounted on the vehicle. A self-driving car, for example, could include a number of external and internal sensors including radar, stereoscopic visible-light cameras, LIDAR, infrared (IR) cameras, global positioning system (GPS), and audio. To perform autonomous navigation, the heterogeneous data collected from sensors are used for learning a number of interrelated but complex tasks such as localization and mapping, scene understanding, movement planning, and driver-state recognition.

One of the biggest challenges for autonomous navigation is the dynamic nature of the operating environment—the system has to adapt and be reactive to weather variations, lighting variability, pedestrians and other traffic, road conditions, and traffic signs, as well as the driver's state. Nevertheless, deep-learning and reinforcement-learning techniques [26] have been instrumental in advancing this field of application with industry players like Uber, Nvidia, Baidu, and Tesla actively involved in the development of commercial self-driving cars.

An important task in autonomous driving is real-time scene understanding. It requires the learning system to recognize objects in the scene, like lanes, traffic signs, pedestrians, and other traffic. It follows that, for each frame of the multimodal video feed, semantic segmentation has to first be performed. Each semantic concept identified in the scene then has to be localized. For such tasks, deep fully convolutional architectures that perform pixel-wise labeling of each frame are often used [27]. For multimodal inputs, a common strategy is to concatenate synchronized frames across the channel dimensions before being input to a fully convolutional neural network (CNN) (this is, in a sense, early fusion) or, alternatively, to train separate modality-wise networks and then fuse at a deeper stage in the multimodal network. We further discuss such fusion strategies in the section "Fusion Structure." Semantic segmentation can be extended to video by using a 3-D variant of fully CNNs. The basic techniques used in self-driving vehicle technology can be extended to other robotic applications such as mobile robots or drone navigation, grasp configuration learning [28], and robotic manipulation [29].

Summary

We have highlighted three major areas where deep multimodal learning approaches have gained a foothold and continue to experience rapid advancements. In addition to the work already highlighted with respect to each of these key areas, we list additional representative work involving deep multimodal learning in Table 3. Several other application areas that involve text, images, and video, e.g., visual question answering (VQA) and image and video annotations, are highlighted in subsequent sections where we discuss specific deep-learning models.

Models

Applying multimodal deep learning to a new problem involves the selection of both an architecture and a learning algorithm. Together, we will call these choices a *model*. A plethora of different deep-learning models have been proposed for multimodal data. While there are several ways that they could be partitioned and organized for review, we have opted to group notable examples according to their learning paradigm, specifically whether they are generative, discriminative, or hybrid models. Our reason for choosing this categorization is that this choice impacts the available architectures and learning algorithms from which to select.

Generative models implicitly or explicitly represent a data distribution, often allowing for new data to be sampled or "generated" through a process, hence their name. Discriminative models, on the other hand, are less ambitious. Rather than modeling distributions, they attempt to model class boundaries. In the supervised learning setup, where we have data X and labels Y, generative models learn the joint probability $P(X, Y)$. In contrast, discriminative models are used for primarily prediction tasks, and these models learn the conditional $P(Y \mid X)$. Yet generative models can still have discriminative properties. An advantage of generative models is that they are much more flexible. For example, $P(X, Y)$ can be sampled in the case of missing modalities during inference.

Discriminative models

Discriminative deep architectures directly model the mapping from inputs to outputs, and the model parameters are learned

Table 3. Diverse applications of multimodal deep learning.

Reference	Year	Modalities	Problem	Fusion Method	Model	Architecture
Ngiam et al. [6]	2011	Audio,video	Speech classification	Intermediate	Generative	Sparse RBM
Srivastava and Salakhutdinov [30]	2012	Image, text	Image annotation	Intermediate	Generative	DBN
Cao et al. [31]	2014	Medical images, textual descriptions	Content-based medical image retrieval	Intermediate	Generative	DBM
Liang et al. [32]	2015	Gene expression, DNA methylation, and drug response	Cancer subtype clustering	Intermediate	Generative	DBM
Valada et al. [15]	2016	Multispectral imagery	Semantic segmentation	Early	Discriminative	FCNN
Simonyan and Zisserman [33]	2014	Image and optical flow	Action recognition	Late	Discriminative	CNN
Kahou et al. [11]	2015	Video, audio	Emotion recognition	Late	Discriminative	CNN, RNN, SVM, and AE
Liu et al. [20]	2015	MRI, PET	Medical diagnosis	Intermediate	Discriminative	Stacked AE, SVM
Poria et al. [34]	2015	Video, audio, text	Sentiment analysis	Intermediate, late	Discriminative	CNN, SVM
Lenz et al. [28]	2015	Intensity, depth video	Robotic grasping	Intermediate	Discriminative	Stacked AE and MLP
Jain et al. [35]	2016	Video features, GPS coordinates, vehicle dynamics	Driver activity anticipation	Intermediate	Discriminative	LSTM

by minimizing some regularized loss function. Such models compose the bulk of the proposed models for multimodal learning, while tasks include classification or recognition for a variety of problem domains.

In addition to the aforementioned active research problems, image captioning and VQA [36], both of which combine natural language processing and high-level scene interpretation by machine-learning algorithms, have garnered active research interest. In deep image captioning, the model is required to generate a textual description of image content, and this could be achieved by using both discriminative techniques [37], [38] and generative approaches [39]. On the other hand, VQA typically requires the model to answer complex questions based on image content, which is a generative task. This problem can also be cast into a discriminative setting (e.g., multiple choice questions) [40]. Recently, Kim et al. [41] extended the highly successful deep residual network model for a multimodal VQA problem. As multiple modalities may have correlations, the authors carefully designed joint residual mappings across modalities and achieved state-of-the-art results for VQA.

Discriminative deep multimodal learning models have also been proposed for human activity recognition. With the cheap availability of RGB-depth (RGB-D) cameras and ubiquity of smartphones with numerous sensors, deep multimodal learning architectures that involve from four to five modalities have been reported. These problems involve temporal data (video, joint motion, audio), and it is essential that spatiotemporal dependencies be learned effectively. To capture temporal structures and relationships, deep multimodal learning approaches typically use temporal components such as LSTMs or hidden Markov models combined with visual representation learning layers like CNNs or 3-D-CNNs [42], [43]. These models have benefited from the combination of CNNs and recurrent layers that can collectively capture spatiotemporal relationships.

There are also instances of work where generative models have been adapted to perform discriminative tasks. For example, a discriminative variant of the RBM [building block of deep belief networks (DBNs) and deep Boltzmann machines] was proposed by Larochelle and Bengio [44]. Other discriminative models have previously been mentioned while discussing application areas in the sections "Human Activity Recognition," "Medical Applications," and "Autonomous Systems." In addition, Table 3 also lists examples of discriminative models for other multimodal problems. While discriminative models excel at the task of classification or regression, they cannot cope when there are missing data or modalities. Discriminative models also require a large set of labeled data, which could be expensive to obtain in certain applications. Next, we review deep generative multimodal models, which offer some advantages, considering the drawbacks of discriminative models, in the context of learning multimodal representations.

Generative models

Deep generative models typically characterize the high-order correlation properties of the observed or visible data for pattern analysis or synthesis purposes. They can also be used to characterize the joint statistical distributions of the visible data and their associated classes. Generative models like DBNs can also be used for classification and regression tasks by exploiting their capability to learn (unsupervised) from unlabeled data and fine-tuned in a discriminative setting using the backpropagation algorithm or by using the learned representation in conjunction with other classifiers such as support vector machines (SVMs).

For multimodal learning problems, generative models are useful in situations where there could be missing modalities during test time or when there is a lack of labeled data. The early works of Ngiam et al. [6] and Srivastava and Salakhutdinov [30] proved that generative models are indeed capable of handling such learning problems. Since then, a number of works have been reported in the literature that specifically deal with using generative deep multimodal networks in cases where there are missing data [31], [45].

While energy-based models based on stacking RBMs have received most of the attention in deep generative multimodal learning, the landscape of generative models is changing. Recently, generative adversarial networks [46], deep directed models trained with variational inference [47], are gaining traction in multi- and unimodal settings [48]–[50].

Hybrid models

While discriminative models are trained to maximize the separation between classes, generative models excel at modeling data distributions. Hybrid models combine both discriminative and generative components in a unified framework. Deng [51] defines hybrid deep architectures as architectures where the goal is discrimination but is assisted (often in a significant way) with the outcomes of generative architectures. For example, the generative component in a hybrid model may learn a deep representation of input modalities and use the discriminative component for classification or regression tasks.

Hybrid models can be divided into three groups as per [52]:

1) *joint methods* that optimize a single objective function to learn a joint representation using the generative and discriminative components
2) *iterative methods* that learn a shared representation using an iterative method such as expectation maximization using representations updated from both generative and discriminative components
3) *staged methods,* where the generative and discriminative components are trained separately in stages.

Representations learned by the generative model in an unsupervised manner can then be used as features for the discriminative component using supervised training.

An example of a joint model is reported in [53], where short-term temporal characteristics and long-range temporal dependencies for audio-video modalities are modeled by combining a conditional RBM temporal generative model for the former and a discriminative component consisting of a conditional random field for the latter. This model also is able to handle missing modalities due to the generative component. Other related hybrid architectures include those of Sachan et al. [54] and Liu et al. [55].

Summary

In this section, we have highlighted multimodal architectures according to their primary learning paradigm. In some sense, deep-learning models can be thought of as building blocks that allow us to "mix and match" different models to create elaborate deep multimodal architectures. While this can be seen as an advantage of deep learning, a common issue is that architecture design has been more an art than a science. Not withstanding, there are numerous hyperparameters associated with each model that have to be carefully fine-tuned, and this process may be possibly even more complicated when dealing with hybrid architectures. Another aspect to be concerned with is the choice of the fusion structure between modalities and their representations. Next, we discuss several choices for multimodal fusion structure and direct our discussion to the attractive notion of optimizing and learning this fusion architecture for improved performance.

Fusion structure

Deep architectures offer the flexibility of implementing multimodal fusion either as early, intermediate, or late fusion. Multimodal fusion approaches predating the advent of deep learning often referred to early fusion as *feature-level fusion* and late fusion as *decision-level fusion*. With deep-learning approaches, however, the idea of feature-level fusion can be extended further to the concept of intermediate fusion.

Early fusion

Early fusion involves the integration of multiple sources of data, at times very disparate, into a single feature vector, before being used as input to a machine-learning algorithm as illustrated in Figure 2(a). The data to be fused are the raw or preprocessed data from the sensor; hence, the terms *data fusion* or *multisensor fusion* are often used.

If data fusion is performed without feature extraction, this could be quite challenging. For instance, the sampling rate between different sensors could vary, or synchronized data from multiple data sources might not be available if one source produces discrete data, while another source provides a continuous data stream.

To alleviate some of the issues related to fusing raw data, higher-level representations can first be extracted from each modality, which could be either handcrafted features or learned representations, as is common in deep learning, before fusing at the feature level. When nonhierarchical features are used, as often the case in handcrafted features, features extracted from multiple modalities can be fused at only one level, prior to being input to the machine-learning algorithm. Since deep learning essentially involves learning hierarchical representations from raw data, this gives rise to intermediate-level fusion.

Most early-fusion models make the simplifying assumption that there is conditional independence between the states of various sources of information. This may not be true in practice, as multiple modalities tend to be highly correlated (for example, video and depth cues). Sebe [56] argues that different streams contain information that is correlated to another stream only at a high level. An excellent example of this can be seen in [57]. This assumption allows the output of each modality to be processed independently of the others.

In its simplest form, early fusion involves concatenation of multimodal features as was implemented by Poria et al. [34]. Early fusion of multimodal data may not fully exploit the complementary nature of the modalities involved and may lead to very large input vectors that may contain redundancies. Typically, dimensionality reduction techniques like PCA are applied to remove these redundancies in the input space. Autoencoders, which are nonlinear generalizations of PCA [58], are popularly used in deep learning to extract a distributed representation from raw data. This idea has been extended to learn a multimodal embedded space with the aim to represent multimodal data within a common feature space [59], [60].

One of the issues faced in early fusion of multimodal data is to determine the time-synchronicity between different data sources. Commonly, these signals are resampled at a common sampling rate. To mitigate this issue, Martínez and Yannakakis [61] proposed several methods (convolution, training, and pooling fusion) to integrate sequences of discrete events with continuous signals.

Late fusion

Late- or decision-level fusion refers to the aggregation of decisions from multiple classifiers, each trained on separate modalities [see Figure 2(b)]. This fusion architecture is often favored because errors from multiple classifiers tend to be uncorrelated and the method is feature independent. Various rules exist to determine how decisions from different classifiers are combined.

These fusion rules could be max-fusion, averaged-fusion, Bayes' rule based, or even learned using a metaclassifier. Decision-level fusion was popular in the early- to mid-2000s, when ensemble classifiers received widespread interest within the machine-learning community.

There have been several works that employ late- or decision-level fusion for deep multimodal learning [33], [43], [62] in addition to some works listed in Table 3. Based on the papers that we have reviewed, we do not find conclusive evidence that late fusion is better than early fusion—the performance is very much problem dependent. Undoubtedly, when input modalities are significantly uncorrelated, of very different dimensionality and sampling rates, it is much simpler to implement a late-fusion approach for multimodal learning problems. An alternative approach, intermediate fusion, offers much more flexibility as to how and when representations learned from multimodal data can be fused.

Intermediate fusion

Neural networks transform raw inputs to higher-level representations by mapping the input through a pipeline of layers. Each layer typically alternates linear and nonlinear operations that scale, shift, and skew its input, producing a new representation of the original data. In the multimodal context, when all of the modalities are transformed into representations, then it becomes amenable to fuse different representations into a single hidden layer and then learn a joint multimodal representation. The majority of work in deep multimodal fusion adopts this intermediate-fusion approach, where a shared representation layer is constructed by merging units with connections coming into this layer from multiple modality-specific paths. Figure 2(c) illustrates a simple intermediate fusion model with three modalities. Representations (features) are learned using different kinds of layers (e.g., 2-D-convolution, 3-D-convolution, or fully connected), and representations are fused using a fusion layer, also known as a *shared representation layer.*

This shared representation layer can be a single shared layer that fuses multiple channels at some depth or could be gradually fused, one or more modalities at a time. A naïve concatenation of features or weights in the shared representation layer may lead to overfitting or the network failing to learn associations between modalities due to distinct underlying distributions. A simple method of improving performance of multimodal fusion is to apply some form of dimensionality reduction like PCA [63] or stacked autoencoders [10] after constructing a shared representation layer (or fusion layer) via simple concatenation of weights from different modalities. This choice of fusing various representations at different depths is perhaps the most powerful and flexible aspect of deep multimodal fusion as opposed to other fusion techniques. The advantage of a flexible fusion scheme can be seen in the

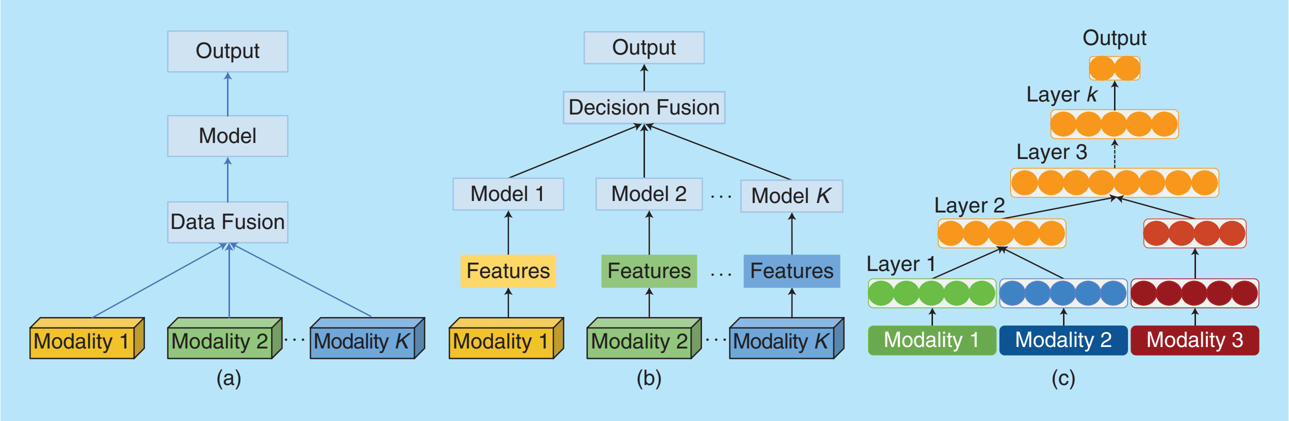

FIGURE 2. An illustration of various fusion models for multimodal learning. (a) Early or data-level fusion, (b) late or decision-level fusion, and (c) intermediate fusion.

work of Karpathy et al. [64], who showed that using a "slow-fusion" model, where learned representations of video streams are gradually fused across multiple fusion layers during training, consistently produced better results for a large-scale video classification problem, as opposed to early-fusion and late-fusion models. Similarly, Neverova et al. [8] empirically showed that implementing a gradual fusion strategy, by first fusing highly correlated modalities, to less correlated ones in a progressive manner (e.g., visual modalities first, then motion capture, then audio), produced state-of-art results for communicative gesture recognition.

Although learning a multimodal representation using the shared representation layer is indeed flexible, many current architectures require careful design in terms of how, when, and which modalities can be fused. In the "Fusion Structure Learning and Optimization" section, we discuss further attempts at optimizing the tedious architecture design process required by multimodal learning.

Multimodal regularization

Deep-learning techniques iteratively optimize a set of model parameters (typically, the weights and biases between each layer) by minimizing a loss function. To improve generalization, one or more regularization strategies are employed, often as an additional term added to the loss function. From a computational perspective, regularization provides stability to the optimization problem leading to algorithmic speed-ups, and, from a statistical point of view, regularization reduces overfitting [65].

In the deep multimodal learning context, an important design consideration is the formulation of cost functions and regularizers that enforce intermodality and intramodality relationships such as information-theoretic regularizers and structured regularization, which we now briefly review.

Information-theoretic regularizers are formulated using measures such as mutual information and variation of information. For example, Sohn et al. [66] proposed a cost function that minimized the variation of information between modalities to learn relationships between modalities. The intuition behind this formulation is that learning to maximize the amount of information that one data modality has about the others would allow generative models to reason about the missing data modality given partial observations. Alternatively, a mutual information term could also be maximized during training [67]. Another information theoretic loss formulation based on the Kulback–Leibler (KL) divergence was proposed by Zhu et al. [68] for a multilabel image annotation problem. In a pretraining stage, they first trained CNNs to learn intermediate representations from each modality using unlabeled data and then, in the fine-tuning stage, used backpropagation to minimize the KL-divergence between the predictive and ground-truth distributions. Finally, to learn the optimal combination of multimodal weights, they adopted the exponentiated online learning algorithm to sequentially find an optimal set of combinational weights.

Taking inspiration from structured feature selection in multitask learning [69], Wu et al. [70] designed a model that uses a trace norm regularization term, which encourages similar modalities to share similar representations for a video classification problem using video and audio modalities.

Cost functions that enforce inter- and intramodality correlations have also been explored by Wang et al. [71]. Their formulation includes a discriminative term, and a correlative term based on canonical correlation analysis. In a follow-up work [72], they proposed a multimodal fusion layer that uses matrix transformations to explicitly enforce a common part to be shared by features of different modalities while retaining modality-specific learning.

Lenz et al. [28] formulated a structured regularization term in the cost function, which allows a model to learn correlated features between multiple input modalities but regularizes the number of modalities used per feature, thereby discouraging the model from learning weak correlations between modalities. Structured regularization essentially applies some form of regularization separately for each set of modality-specific weights. They considered several variants of structure regularization for a multimodal robotic grasping task. One that worked well in their case incorporated the L_0 norm on top of the max-norm penalty

$$f(W) = \sum_{j=1}^{K} \sum_{r=1} \mathbb{I}\left\{\left(\max_i S_{r,i}\left|W_{i,j}\right|\right) > 0\right\},$$

where $S_{r,i}$ is one if feature i belongs to group r and is otherwise zero. S is a binary modality matrix of size $R \times N$, where each element $S_{r,i}$ indicates the membership of a visible unit, x_i, in a particular modality, r. $\mathbb{I}$ is an indicator function, which takes a value of one if its argument is true and is otherwise zero.

In some problems, temporal context can play an important role, for example, driver activity anticipation. Unlike human activity recognition, where complete temporal context is available, in driver activity anticipation, the machine-learning system must predict using only partial context within a short span of time before the event occurs. To solve this problem, Jain et al. [35] incorporated a temporal term that grows exponentially in time into their cost function for a multimodal RNN with LSTM units. This encourages the model to fix mistakes as early as it can.

Multimodal-aware regularizers have resulted in marginal to notable improvements in model performance. Despite including these multimodal regularization strategies, the deep-learning architectures discussed in this section have input modalities merging into a single fusion layer. A possible extension could be to investigate a gradual fusion model that takes advantage of these regularization strategies.

Fusion structure learning and optimization

Most multimodal deep-learning architectures proposed to date are meticulously handcrafted. While many models adopt a single fusion layer (shared representation layer), several stand-out works [8], [64] implemented a gradual fusion strategy. The choice of which modality is fused, and at which depth of representation, is usually based on intuition (for example, fusing similar modalities early, and then fusing disparate modalities at a deeper layer). When more than two modalities are involved, also depending upon the nature of the modalities being used in the problem, choosing an optimal fusion architecture may be more challenging. A natural progression would be to search for an optimal multimodal fusion architecture by casting this as a model search or structure learning problem.

Neural network structure optimization for unimodal problems has long been investigated by machine-learning researchers. These mainly involved determining the optimal number of neurons and number of layers in a network. There is a tradeoff between good generalization ability of the network, and the number of parameters and availability of training data. Too large a network might perform well or overfit, depending if it is trained with sufficiently large training data, while too small a network, might underfit and may result in poor generalization.

A common approach is to adopt a bottom-up constructive approach. The basic idea proposed by Elman [73] is to start with a relatively small network and add hidden units or layers incrementally until the best performing architecture is found. More recently, and in the large-scale setting, Chen et al. [74] gradually added depth and width to an inception-style [75] network by knowledge transfer between one neural network to another.

Pruning algorithms [76] address the same problem from a top-down approach. Recent approaches for DNNs include the works of Feng and Darrel [77], who proposed an evolving grow-and-prune algorithm that optimizes the structure of an Indian buffet process-CNN model, and Yang et al. [78], who introduced network pruning for large, diverse data sets based on sparse representations.

Genetic algorithm (GA)-based structure optimization of neural networks was one of the earliest metaheuristic search algorithms used for neural network structure search and optimization [79]. In the early 2000s, an algorithm called *Neuro Evolution of Augmenting Topologies* (*NEAT*) [80] that also used GAs to evolve increasingly complex neural network architectures received much attention. More recently, Shinozaki and Watanabe [81] applied GAs and a covariance matrix evolution strategy to optimize the structure of a DNN, parameterizing the structure of the DNN as a simple binary vector based on a directed acyclic graph representation. As the GA search space can be very large, and each model evaluation in the search space is expensive, a parallel search using a large GPU cluster was used to speed up the process.

These neural network structural search and optimization techniques can readily be extended to the multimodal setting if a suitable representation of the network architecture is devised and provided that the cost of training and testing multiple architectures during the search process is not prohibitively expensive. With data set sizes approaching gigabytes, and even terabyte levels, and deep network architectures involving millions of parameters and multiple modalities, search and optimization of multimodal fusion structure can be prohibitively expensive unless some parallel search procedure is implemented or an efficient optimization algorithm is used. While Bayesian optimization (BO) [82] has been a popular choice for hyperparameter optimization, it has been recently used for multimodal fusion architecture optimization [83]. Architecture optimization was cast as a discrete optimization problem by searching a space of all possible multimodal fusion architectures using a Gaussian process-based BO. A novel graph-induced kernel was proposed to quantify the distance between different architectures in the search space.

Reinforcement learning [84] has also been used for deep neural architecture search [85]. This work proposed a novel method of using an RNN to generate variable-length model descriptions of neural networks. The RNN was trained with reinforcement learning to maximize the expected accuracy of the generated architectures on a validation set.

A number of recent works have approached structure learning as a means of regularization, or capacity control, in a network. By pruning the network in a stochastic manner, stochastic regularization methods can be considered as a kind of ensemble that improves generalization via model averaging. Kulkarni et al. [86] implemented a method of learning the structure of DNNs via deterministic regularization. They insert, between each pair of fully connected layers, a sparse diagonal matrix whose entries are l_1 penalized. This implicitly defines the size of the effective weight matrices at each layer. The approach has a similar effect to Dropout [87]. Blockout [88] can perform simultaneous regularization and model selection through a clever technique that stochastically assigns hidden units to "clusters," forming block-structured weight matrices. In addition, by averaging the outputs of multiple stochastic inference passes (which can be viewed as a case of ensemble classifiers), results better than ResNets were achieved. This architecture effectively implements a late fusion of multiple architectures to achieve better results.

Stochastic regularization has been extended to the multimodal setting by Neverova et al. [8] and, more recently, by Li et al. [89]. In the latter work, the authors show that, when the intermodality correlation is high, an early-fusion approach (whose fusion structure was learned by the network) produced better results, while a late-fusion approach worked better when the input modalities are less correlated. This concurred with the empirical choice made by the former.

In this section, we have covered a number of recent works that use either stochastic regularization or optimization resulting in deep multimodal fusion architectures that perform at par with or better than meticulously designed ones. While feature engineering has been largely solved by deep representation learning, the next logical step would be to do away with meticulous engineering of deep architectures and pursue techniques that achieve this automatically.

Data sets

To facilitate research in multimodal learning, a number of data sets have been released to the public. We note that the majority of these data sets typically involve person-centric visual understanding, with variants including emotion recognition, group behavior analysis, etc. Table 2 lists a number of such data sets, the modalities involved, and the problem domain. While this list is not exhaustive, we cover more recent data sets (many of which were released in the past three years) that are available for multimodal research. While most data sets include at least two modalities (images and text, for example) or up to four (RGB-D, audio, and skeletal pose), some data sets, for example, H-MOG [12], include up to nine different modalities. For the interested reader, Firman [90] presents an extensive survey of 102 RGB-D data sets. Autonomous driving and driver assistance systems (using driver behavior prediction) are being pursued as a popular research topic in deep learning. Such data sets are not only highly multimodal [91], with data from up to six individual sensors, but also very large—hours of data available. The Oxford RobotCar [92] data set, for example, contains more than 23 TB of year-long driving data in various weather conditions.

We note that there are relatively fewer multimodal medical data sets available, possibly due to the cost and ethical and privacy concerns. Most medical data sets also tend to be much smaller, involving between ten and 50 subjects and also suffer from high class imbalances (for example, it is much more common to have normal cases in comparison to abnormal cases). Medical informatics and imaging studies rely heavily on multimodal information, and this can be leveraged to improve computer-aided diagnosis. Efforts to gather and make such data sets publicly available are encouraged.

Conclusions and future directions

In this article, we have reviewed recent advancements in deep multimodal learning. It is undeniable that the incorporation of multiple modalities into the learning problem almost always results in much better performance for a wide range of problems. From a fusion perspective, we see that techniques in deep multimodal learning can be classified into early- and late-fusion approaches and that deep-learning methods facilitate a flexible intermediate-fusion approach, which not only makes it simpler to fuse modality-wise representations and learn a joint representation but also allows multimodal fusion at various depths in the architecture. Although deep learning has, in many cases, reduced the need for feature engineering, deep-learning architectures still involve a great deal of manual design, and experimenters may not have explored the full space of possible fusion architectures. It is only natural that researchers should extend the notion of learning to architectures in an effort to have a truly generic learning method, which can be adapted, with minimal or no human intervention, to a specific task.

We reviewed several options for learning an optimal architecture. This includes stochastic regularization, casting architecture optimization as a hyperparameter optimization problem using, for example, BO, and incremental online reinforcement learning. This is, in our opinion, the most exciting area of research for deep multimodal learning. Architecture learning can be extremely compute-intensive, so researchers should take advantage of advances in hardware acceleration and distributed deep learning.

We have also identified several application domains that are gaining the most attention in deep multimodal learning. This includes RGB-D and data from the multitude of sensors on mobile phones that have been used for a range of problems involving multimodal data such as human activity recognition and their variants. We foresee that this area will gain more attention in the coming years for novel applications, which will profoundly impact our daily lives. Another important area highlighted is medical research, which involves numerous modalities of data, some of which are very difficult to interpret without human experts in the loop. With the medical community opening up to the rise of artificial intelligence-assisted diagnosis, we will see more significant progress being made in this domain. Finally, two more application areas that are gaining the attention of deep-learning researchers involve autonomous vehicles or robotics and multimedia applications, for example, video transcription, image captioning, etc. Novel applications like online chatbots that use multimodal inputs, like images, and text or recommender systems that utilize multimodal data may become widespread in the near future.

We conclude by acknowledging that this is very much a fast-evolving field, and, at the rate of the amount of new research being published, many new innovations in deep multimodal learning architectures are bound to be presented. We have tried not to provide specific suggestions to architecture design, as we found many problems require application-specific considerations. Regardless, we feel this is a timely publication as the directions of future research that we have highlighted, hopefully, can act as a guide toward a more organized effort into advancing the research field.

Authors

Dhanesh Ramachandram (dramacha@uoguelph.ca) received his B.Tech degree in industrial technology and his Ph.D. degree in computer vision and robotics from the Universiti Sains Malaysia in 1997 and 2003, respectively, where he was formerly an associate professor. He is a researcher at the University of Guelph, Ontario, Canada, and a Senior Member of the IEEE. He is interested in deep learning for computer vision, medical imaging, and multimodal problems.

Graham W. Taylor (gwtaylor@uoguelph.ca) received his received his bachelor's and master's degrees in applied science from the University of Waterloo, Canada, in 2003 and 2004, respectively. He received his Ph.D. degree in computer science from the University of Toronto, Canada, in 2009, where his thesis coadvisors were Geoffrey Hinton and Sam Roweis. He is an associate professor at the University of Guelph, Ontario, Canada, a member of the Vector Institute for Artificial Intelligence, and a Canadian Institute for Advanced Research Azrieli Global Scholar. He is interested in statistical machine learning and biologically inspired computer vision, with an emphasis on unsupervised learning and time-series analysis.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] D. Lahat, T. Adali, and C. Jutten, “Multimodal data fusion: An overview of methods, challenges, and prospects,” *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[3] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: A survey,” *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[4] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, “Multisensor data fusion: A review of the state-of-the-art,” *Inform. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[5] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proc. 28th Int. Conf. Machine Learning (ICML-11)*, 2011, pp. 689–696.

[7] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep Boltzmann machines,” in *Proc. Advances in Neural Inform. Processing Syst*., 2012, pp. 2222–2230.

[8] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, “ModDrop: Adaptive multi-modal gesture recognition,” *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 38, no. 8, pp. 1692–1706, 2016.

[9] S. S. Mukherjee and N. M. Robertson, “Deep head pose: Gaze-direction estimation in multimodal video,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.

[10] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[11] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, et al., “EmoNets: Multimodal deep learning approaches for emotion recognition in video,” *J. Multimedia User Interfaces*, vol. 10, no. 2, pp. 99–111, 2015.

[12] Z. Sitová, J. Šeděnka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K. S. Balagani, “HMOG: New behavioral biometric features for continuous authentication of smartphone users,” *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 877–892, 2016.

[13] V. Radu, N. D. Lane, S. Bhattacharya, C. Mascolo, M. K. Marina, and F. Kawsar, “Toward multimodal deep learning for activity recognition on mobile devices,” in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 185–188.

[14] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, “Modeep: A deep learning framework using motion features for human pose estimation,” in *Proc. Asian Conf. Computer Vision*, 2014, pp. 302–315.

[15] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, “Deep multispectral semantic scene understanding of forested environments using multimodal fusion,” in *Proc. Int. Symp. Experimental Robotics (ISER 2016)*, 2016, pp. 465–477.

[16] D. Shen, G. Wu, and H.-I. Suk, “Deep learning in medical image analysis,” *Annu. Review Biomedical Eng*., vol. 19, pp. 221–248, 2017.

[17] R. Kiros, K. Popuri, D. Cobzas, and M. Jagersand, “Stacked multiscale feature learning for domain independent medical image segmentation,” in *Proc. Int. Workshop on Mach. Learning in Medical Imaging*, 2014, pp. 25–32.

[18] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, “Online multimodal deep similarity learning with application to image retrieval,” in *Proc. 21st ACM Int. Conf. Multimedia*. 2013, pp. 153–162.

[19] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, “A deep metric for multimodal registration,” in *Proc. Int. Conf. Medical Image Computer and Computer-Assisted Intervention*, 2016, pp. 10–18.

[20] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, et al., “Multimodal neuroimaging neature learning for multiclass diagnosis of Alzheimer’s disease,” *IEEE Trans. Biomed. Eng*., vol. 62, no. 4, pp. 1132–1140, 2015.

[21] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, “Applications of deep learning in biomedicine,” *Molecular Pharmaceutics*, vol. 13, no. 5, pp. 1445–1454, 2016.

[22] Y. Guo, G. Wu, L. A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen, “Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features,” in *Proc. Int. Conf. Medical Image Computer and Computer-Assisted Intervention*, 2014, p. 308.

[23] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Trans. Med. Imag*., vol. 35, no. 5, pp. 1299–1312, 2016.

[24] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, “Interpretable deep neural networks for single-trial eeg classification,” *J. Neuroscience Methods*, vol. 274, pp. 141–145, Dec. 2016.

[25] S. G. Kim, N. Theera-Ampornpunt, C.-H. Fang, M. Harwani, A. Grama, and S. Chaterji, “Opening up the blackbox: an interpretable deep neural network-based classifier for cell-type specific enhancer predictions,” *BMC Syst. Biology*, vol. 10, no. 2, p. 54, 2016.

[26] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “Deepdriving: Learning affordance for direct perception in autonomous driving,” in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 2722–2730.

[27] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[28] I. Lenz, H. Lee, and A. Saxena, “Deep learning for detecting robotic grasps,” *Int. J. Robotics Res*., vol. 34, no. 4-5, pp. 705–724, 2015.

[29] S. Gu, E. Holly, T. Lillicrap, and S. Levine. (2016). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. *arXiv*. [Online]. Available: https://arxiv.org/abs/1610.00633

[30] N. Srivastava and R. Salakhutdinov, “Learning representations for multimodal data with deep belief nets,” presented at *Proc. 29th Int. Conf. Machine Learning (Workshop)*, 2012.

[31] Y. Cao, S. Steffey, J. He, D. Xiao, C. Tao, P. Chen, and H. Müller, “Medical image retrieval: A multimodal approach,” *Cancer Informatics*, vol. 13, no. Suppl 3, p. 125, 2014.

[32] M. Liang, Z. Li, T. Chen, and J. Zeng, “Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach,” *IEEE/ACM Trans. Comput. Biol. Bioinf*., vol. 12, no. 4, pp. 928–937, 2015.

[33] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[34] S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” in *Proc. Conf. Empirical Methods on Natural Language Processing*, 2015, pp. 2539–2544.

[35] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, “Recurrent neural networks for driver activity anticipation via sensory-fusion architecture,” in *Proc. 2016 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2016, pp. 3118–3125.

[36] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proc. Int. Conf. Computer Vision (ICCV)*, 2015, pp. 2425–2433.

[37] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[38] A. Karpathy, A. Joulin, and F. F. F. Li, “Deep fragment embeddings for bidirectional image sentence mapping,” in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 1889–1897.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[40] M. Ren, R. Kiros, and R. Zemel, “Exploring models and data for image question answering,” in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 2953–2961.

[41] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. (2016). Multimodal residual learning for visual QA. *arXiv*. [Online]. Available: https://arxiv.org/abs/1606.01455

[42] F. J. Ordóñez and D. Roggen, “Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, p. 115, 2016.

[43] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, “Deep dynamic neural networks for multimodal gesture segmentation and recognition,” *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 38, no. 8, pp. 1583–1597, 2016.

[44] H. Larochelle and Y. Bengio, “Classification using discriminative restricted Boltzmann machines,” in *Proc. 25th Int. Conf. Machine Learning*, 2008, pp. 536–543.

[45] Y. Huang, W. Wang, and L. Wang, “Unconstrained multimodal multi-label learning,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, 2015.

[46] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[47] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

[48] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, 2016, pp. 1060–1069.

[49] M. Suzuki, K. Nakayama, and Y. Matsuo. (2016). Joint multimodal learning with deep generative models. *arXiv*. [Online]. Available: https://arxiv.org/abs/1611.01891

[50] G. Pandey and A. Dukkipati. (2016). Variational methods for conditional multimodal deep learning. *arXiv*. [Online]. Available: https://arxiv.org/abs/1603.01801

[51] L. Deng, “A tutorial survey of architectures, algorithms, and applications for deep learning,” *APSIPA Trans. Signal and Inform. Processing*, vol. 3, pp. 1–29, 2014.

[52] M. R. Amer, T. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, and S. Chai, “Deep multimodal fusion: A hybrid approach,” *Int. J. Comput. Vision*, pp. 1–17, 2017. DOI: 10.1007/s11263-017-0997-7.

[53] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, “Multimodal fusion using dynamic hybrid models,” in *Proc. IEEE 2014 Applications of Computer Vision Winter Conf.*, 2014, pp. 556–563.

[54] D. S. Sachan, U. Tekwani, and A. Sethi, “Sports video classification from multimodal information using deep neural networks,” in *Proc. 2013 Association for the Advancement of Artificial Intelligence Fall Symp.*, 2013, pp. 102–107.

[55] Y. Liu, X. Feng, and Z. Zhou, “Multimodal video classification with stacked contractive autoencoders,” *Signal Processing*, vol. 120, pp. 761–766, Mar. 2016.

[56] N. Sebe, *Machine Learning in Computer Vision*, vol. 29. Dordrecht, The Netherlands: Springer, 2005.

[57] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, *Ambient Sound Provides Supervision for Visual Learning*. Cham, Switzerland: Springer International Publishing, 2016, pp. 801–816.

[58] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[59] D. Wang, P. Cui, M. Ou, and W. Zhu, “Learning compact hash codes for multimodal representations using orthogonal deep structure,” *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1404–1416, 2015.

[60] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, “Multimodal similarity-preserving hashing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, 2014.

[61] H. P. Martínez and G. N. Yannakakis, “Deep multimodal fusion,” in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 34–41.

[62] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, et al., “Combining modality specific deep neural networks for emotion recognition in video,” in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 543–550.

[63] D. Yi, Z. Lei, and S. Z. Li, “Shared representation learning for heterogenous face recognition,” in *Proc. Automatic Face and Gesture Recognition 11th IEEE Int. Conf. Workshops*, 2015, pp. 1–7.

[64] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[65] M. J. Wainwright, “Structured regularizers for high-dimensional problems: Statistical and computational issues,” *Annu. Rev. Statistics Application*, vol. 1, pp. 233–253, Apr. 2014.

[66] K. Sohn, W. Shang, and H. Lee, “Improved multimodal deep learning with variation of information,” in *Proc. Advances in Neural Information Processing Systems.*, 2014, pp. 2141–2149.

[67] J. J.-Y. Wang, Y. Wang, S. Zhao, and X. Gao, “Maximum mutual information regularized classification,” *Eng. Applicat. Artificial Intell.*, vol. 37, pp. 1–8, Jan. 2015.

[68] S. Zhu, X. Li, and S. Shen, “Multimodal deep network learning-based image annotation,” *IET Electron. Lett.*, vol. 51, no. 12, pp. 905–906, 2015.

[69] H. Fei and J. Huan, “Structured feature selection and task relationship inference for multi-task learning,” *Knowledge and Inform. Syst.*, vol. 35, no. 2, pp. 345–364, 2013.

[70] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, “Exploring inter-feature and interclass relationships with deep neural networks for video classification,” in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 167–176.

[71] A. Wang, J. Lu, J. Cai, T. J. Cham, and G. Wang, “Large-margin multi-modal deep learning for RGB-D object recognition,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1887–1898, Nov. 2015.

[72] A. Wang, J. Cai, J. Lu, and T.-J. Cham, “MMSS: Multi-modal sharable and specific feature learning for RGB-D object recognition,” in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1125–1133.

[73] J. Elman, “Learning and development in neural networks: The importance of starting small,” *Cognition,* vol. 48, no. 1, pp. 71–99, 1993.

[74] T. Chen, I. Goodfellow, and J. Shlens. (2015). Net2Net: Accelerating learning via knowledge transfer. *arXiv*. [Online]. Available: https://arxiv.org/abs/1511.05641

[75] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[76] R. Reed, “Pruning algorithms—A survey,” *IEEE Trans. Neural Netw.*, vol. 4, no. 5, pp. 740–747, 1993.

[77] J. Feng and T. Darrel, “Learning the structure of deep convolutional networks,” in *Proc. Int. Conf. Computer Vision*, 2015, pp. 2749–2757.

[78] J. Yang, J. Ma, M. Berryman, and P. Perez, “A structure optimization algorithm of neural networks for large-scale data sets,” in *Proc. 2014 IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, 2014, pp. 956–961.

[79] D. Whitley, T. Starkweather, and C. Bogart, “Genetic algorithms and neural networks: Optimizing connections and connectivity,” *Parallel Comput.*, vol. 14, no. 3, pp. 347–361, 1990.

[80] K. O. Stanley and R. Miikkulainen, “Efficient evolution of neural network topologies,” in *Proc. Congr. Evolutionary Computation (CEC02)*, 2002, pp. 1757–1762.

[81] T. Shinozaki and S. Watanabe, “Structure discovery of deep neural network based on evolutionary algorithms,” in *Proc. 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4979–4983.

[82] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, “Taking the human out of the loop: A review of Bayesian optimization,” *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[83] D. Ramachandram, M. Lisicki, T. Shields, M. Amer, and G. Taylor, “Structure optimization for deep multimodal fusion networks using graph-induced kernels,” in *Proc. 25th European Symp. Artificial Neural Networks, Computational Intelligence, and Machine Learning (ESANN)*, Bruges, Belgium, 2017, pp. 11–16.

[84] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, vol. 1. Cambridge, MA: MIT Press, 1998.

[85] B. Zoph and Q. V. Le. (2016). Neural architecture search with reinforcement learning. *arXiv*. [Online]. Available: https://arxiv.org/abs/1611.01578

[86] P. Kulkarni, J. Zepeda, F. Jurie, P. Pérez, and L. Chevallier, “Learning the structure of deep architectures using L1 regularization,” in *Proc. British Machine Vision Conf.*, 2015, pp. 23.1–23.11.

[87] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learning Res.*, vol. 15, no. 1, pp. 1929–1958, 1 Jan. 2014.

[88] C. Murdock, Z. Li, H. Zhou, and T. Duerig, “Blockout: Dynamic model selection for hierarchical deep networks,” in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2583–2591.

[89] F. Li, N. Neverova, C. Wolf, and G. Taylor, “Modout: Learning multi-modal architectures by stochastic regularization,” in *Proc. 2017 IEEE Conf. Automatic Face and Gesture Recognition*, 2017, pp. 422–429.

[90] M. Firman, “RGBD data sets: Past, present and future,” in *Proc. CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling, and Analysis*, 2016.

[91] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The Kitti data set,” *Int. J. Robotics Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[92] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 Year, 1000km: The Oxford RobotCar data set,” *Int. J. Robotics Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[93] C. Chen, R. Jafari, and N. Kehtarnavaz, “UTD-MHAD: A multimodal data set for human action recognition utilizing a depth camera and a wearable inertial sensor,” in *Proc. 2015 IEEE Int. Conf Image Processing (ICIP),* 2015, pp. 168–172.

[94] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, et al., “Chalearn looking at people challenge 2014: Data set and results,” in *Proc. Workshop at the European Conf. Computer Vision*, 2014, pp. 459–473.

[95] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Berkeley MHAD: A comprehensive multimodal human action database,” in *Proc. 2013 IEEE Workshop on Applications of Computer Vision*, 2013, pp. 53–60.

[96] A. Pablo, Y. Mollard, F. Golemo, A. C. Murillo, M. Lopes, and J. Civera, “A multimodal human-robot interaction data set,” in *Proc. Neural Information Processing Systems*, 2016, pp. 1–5.

[97] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the Recola multimodal corpus of remote collaborative and affective interactions,” in *Proc. Automatic Face and Gesture Recognition 10th IEEE Int. Conf. Workshops*, 2013, pp. 1–8.

[98] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, “Design, implementation and validation of a novel open framework for agile development of mobile health applications,” *Biomedical Eng. Online*, vol. 14, no. 2, p. S6, 2015. [Online]. Available: https://doi.org/10.1186/1475-925X-14-S2-S6

[99] J. Mao, J. Xu, K. Jing, and A. L. Yuille, “Training and evaluating multimodal word embeddings with large-scale web annotated images,” in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 442–450.

[100] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. (2017). Gated multimodal units for information fusion. *arXiv*. [Online]. Available: https://arxiv.org/abs/1702.01992

[101] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, “Exploiting feature and class relationships in video categorization with regularized deep neural networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. doi: https://doi.org/10.1109/TPAMI.2017.2670560

[102] R. Min, N. Kose, and J.-L. Dugelay, “KinectFaceDB: A kinect database for face recognition,” *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

[103] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, et al., “The multimodal brain tumor image segmentation benchmark (BRATS),” *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, 2015.