Neurocomputing 417 (2020) 302–321

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey on Deep Neural Networks in Speech and Vision Systems

M. Alam [a,1], M.D. Samad [b], L. Vidyaratne [a], A. Glandon [a], K.M. Iftekharuddin [a,*]

[a] Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529, USA
[b] Department of Computer Science, Tennessee State University, Nashville, TN 37209, USA

ARTICLE INFO

Article history:
Received 6 June 2019
Revised 14 April 2020
Accepted 8 July 2020
Available online 26 July 2020
Communicated by Tie-Yan Liu

Keywords:
Vision processing
Speech recognition
Natural language processing
Computational intelligence
Deep learning
Computer vision
Hardware constraints
Embedded systems
Convolutional neural networks
Deep autoencoders
Generative neural networks

ABSTRACT

This survey presents a review of state-of-the-art deep neural network architectures, algorithms, and systems in speech and vision applications. Recent advances in deep artificial neural network algorithms and architectures have spurred rapid innovation and development of intelligent speech and vision systems. With availability of vast amounts of sensor data and cloud computing for processing and training of deep neural networks, and with increased sophistication in mobile and embedded technology, the next-generation intelligent systems are poised to revolutionize personal and commercial computing. This survey begins by providing background and evolution of some of the most successful deep learning models for intelligent speech and vision systems to date. An overview of large-scale industrial research and development efforts is provided to emphasize future trends and prospects of intelligent speech and vision systems. Robust and efficient intelligent systems demand low-latency and high fidelity in resource-constrained hardware platforms such as mobile devices, robots, and automobiles. Therefore, this survey also provides a summary of key challenges and recent successes in running deep neural networks on hardware-restricted platforms, i.e. within limited memory, battery life, and processing capabilities. Finally, emerging applications of speech and vision across disciplines such as affective computing, intelligent transportation, and precision medicine are discussed. To our knowledge, this paper provides one of the most comprehensive surveys on the latest developments in intelligent speech and vision applications from the perspectives of both software and hardware systems. Many of these emerging technologies using deep neural networks show tremendous promise to revolutionize research and development for future speech and vision systems.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

There has been a massive accumulation of human-centric data to an unprecedented scale over the last two decades. This data explosion coupled with rapid growth in computing power has rejuvenated the field of neural networks and sophisticated intelligent systems (IS). In the past, neural networks have mostly been limited to the application of industrial control and robotics. However, recent advancements in neural networks have led to successful applications of IS in almost every aspect of human life with the introduction of intelligent transportation [1–10], intelligent diagnosis and health monitoring for precision medicine [11–14], robotics and automation in home appliances [15], virtual online assistance [16], e-marketing [17], and weather forecasting and natural disasters monitoring [18] among others. The widespread success of IS technology has redefined and augmented human ability to communicate and comprehend the world by innovating 'smart' physical systems. A 'smart' physical system is designed to interpret, act, and collaborate with complex multimodal human senses such as vision, touch, speech, smell, gestures, or hearing. A large body of smart physical systems have been developed targeting two primary senses used in human communication: speech and vision.

The advancement in speech and vision processing systems has enabled tremendous research and development in the areas of human–computer interactions [19], biometric applications [20,21], security and surveillance [22], and most recently in computational behavioral analysis [23–27]. While traditional machine learning and evolutionary computations have enriched IS to solve complex pattern recognition problems over many decades, these techniques have limitations in their ability to process natural data or images in raw data formats. A number of computational steps are used to extract representative features from raw data or images prior to applying machine learning models. This intermediate representa-

---

[*] Corresponding author.

E-mail addresses: malam001@odu.edu (M. Alam), msamad@tnstate.edu (M.D. Samad), lvidy001@odu.edu (L. Vidyaratne), aglan001@odu.edu (A. Glandon), kiftekha@odu.edu (K.M. Iftekharuddin).

[1] Current affiliation: Research and Development, Hitachi America Ltd.,Santa Clara, CA 95054, USA.

https://doi.org/10.1016/j.neucom.2020.07.053
0925-2312/© 2020 Elsevier B.V. All rights reserved.

tion of raw data, known as 'hand-engineered' features, requires domain expertise and human interpretation of physical patterns such as texture, shape, geometry, etc. There are three major problems with 'hand-engineered' features that impede major progress in IS. First, the choice of 'hand-engineered' features is application dependent and involves human interpretation and evaluation. Second, 'hand-engineered' features are extracted from each sample in a standalone manner without the knowledge of inevitable noise and variations in data. Third, 'hand-engineered' features may perform excellently with some inputs but may completely fail to extract quality features in other types of input data. This can lead to high variability in speech and vision recognition performance.

A solution to the limitations of 'hand-engineered' features has emerged through mimicking functions of biological neurons in artificial neural networks (ANN). The potential of ANNs is recently exploited with access to large trainable datasets, efficient learning algorithms, and powerful computational resources. Advancements of ANN over the last decade have led to deep learning [28,29] that has revolutionized several application domains, including computer vision, speech analysis, biomedical image processing, and online market analyses. The rapid success of deep learning over traditional machine learning may be attributed to three factors. First, deep learning offers end-to-end trainable architectures that integrate feature extraction, dimensionality reduction, and final classification. These steps are otherwise treated as standalone sub-systems in conventional machine learning, which may result in suboptimal pattern recognition performance. Second, target-specific and informative features may be learned from both input examples and classification targets without resorting to application-specific feature extractors. Third, deep learning models are highly flexible in capturing complex nonlinear relationships between inputs and output targets at a level that is far beyond the capacity of 'hand-engineered' features.

The remainder of this article is organized as follows. Section 2 discusses deep learning architectures that have been recently introduced to solve contemporary challenges in speech and vision domain. Section 3 provides a comprehensive discussion of real-world and commercial application cases for the technology. Section 4 discusses state-of-the-art results in implementing these sophisticated algorithms in resource-constrained hardware environments. This section also highlights prospects of 'smart' applications in mobile devices. Section 5 discusses several successful and emerging applications of neural networks in state-of-the-art IS. Section 6 elaborates potential developments and challenges in the future for IS. Finally, Section 7 concludes with a summary of the key observations in this article.

2. Design and architecture of neural networks for deep learning

An ANN consists of multiple levels of nonlinear modules arranged hierarchically in layers. This design is inspired by the hierarchical information processing observed in the primate visual system [30,31]. Such hierarchical arrangements enable deep models to learn meaningful features at different levels of abstraction. Several successful hierarchical ANNs known as deep neural networks (DNNs) are proposed in the literature [32]. Few examples include convolutional neural networks [33], deep belief networks [1], stacked autoencoders [34], generative adversarial networks [35], variational autoencoders [36], flow models [37], recurrent neural networks [38], and attention based models [39]. These models extract both simple and complex features similar to the ones witnessed in the hierarchical regions of the primate vision system. Consequently, the models show excellent performance in solving several computer vision tasks, especially complex object recognition [33]. Cichy et al. [30] show that DNN models mimic biological brain function. The results from their object recognition experiment suggest a close relationship between the processing stages in a DNN and the processing scheme observed in the human brain. In the next few sections, we discuss the most popular DNN models and their recent evolutions in various speech and vision applications.

2.1. Convolutional neural networks

One of the first hierarchical models, known as convolutional neural network (CNN/ConvNet) [33,40], learns hierarchical image patterns at multiple layers using a series of 2D convolutional operations. CNNs are designed to process multidimensional data structured in the form of multiple arrays or tensors. For example, a 2D color image has three color channels represented by three 2D arrays. Typically, CNNs process input data using three basic ideas: local connectivity, shared weights, and pooling that are arranged in a series of connected layers. A simplified CNN architecture is shown in Fig. 1. The first few layers are convolutional and pooling layers. The convolutional operation processes parts of the input data in small localities to take advantage of local data dependency within a signal. The convolutional layers gradually yield more highly abstract representations of the input data in deeper layers of the network. The convolution operation in CNN is also repeated to maximize the use of patterns in the input data.

While convolutional layers detect local conjunctions of features from the previous layer, the role of the pooling layer is to aggregate local features into a more global representation. Pooling is performed by sliding a non-overlapping window over the output of the convolutional layer to obtain a "pooled" value for each window. The pooled value is typically the maximum or average value over each window. The maximum value pooling helps a network become robust to small shifts and distortions in input data. The convolutional layer ends by vectorizing the multidimensional data prior to feeding them into fully connected neural networks that perform classification using highly abstracted features from previous layers. The training of all weights in the CNN architecture, including the image filters and fully connected network weights, is performed by applying a regular backpropagation algorithm commonly known as gradient-descent optimization.

2.2. Deep generative models and autoencoders

The hierarchical model of CNN is designed to efficiently learn target-specific features from raw images and videos for vision related applications. However, the major breakthrough of hierarchical models is the introduction of the 'greedy layer-wise' training algorithm for deep belief networks (DBNs) proposed by Hinton et al. [28]. A DBN is built in a layer-by-layer fashion by training each learning module known as the restricted Boltzmann machine (RBM) [41]. RBMs are composed of a visible and a hidden layer. The visible layer represents raw data in a less abstract form. The hidden layer is trained to represent more abstract features by capturing correlations in the visible layer data [41]. Fig. 2 (a) shows a standard architecture of a DBN. DBNs are considered hybrid networks that do not support direct end-to-end learning. Consequently, a more efficient architecture, known as deep Boltzmann machine (DBM) [42], has been introduced. Similar to DBNs, DBMs are structured by stacking layers of RBMs. However, unlike DBNs, the inference procedure of DBMs is bidirectional, allowing them to learn in the presence of more ambiguous and challenging data.

The introduction of DBMs has led to the development of the stacked autoencoder (SAE) [34,43], which is also formed by stacking multiple layers. Unlike DBNs, SAEs utilize autoencoders (AE) [44] as the basic learning module. An AE is trained to learn a copy of the input at its output. In doing so, the hidden layer learns an abstract representation of inputs in a compressed form that is

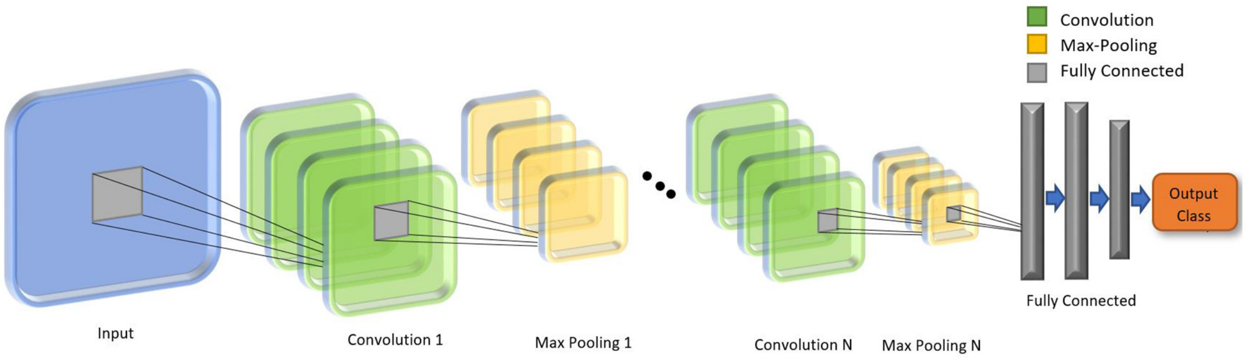
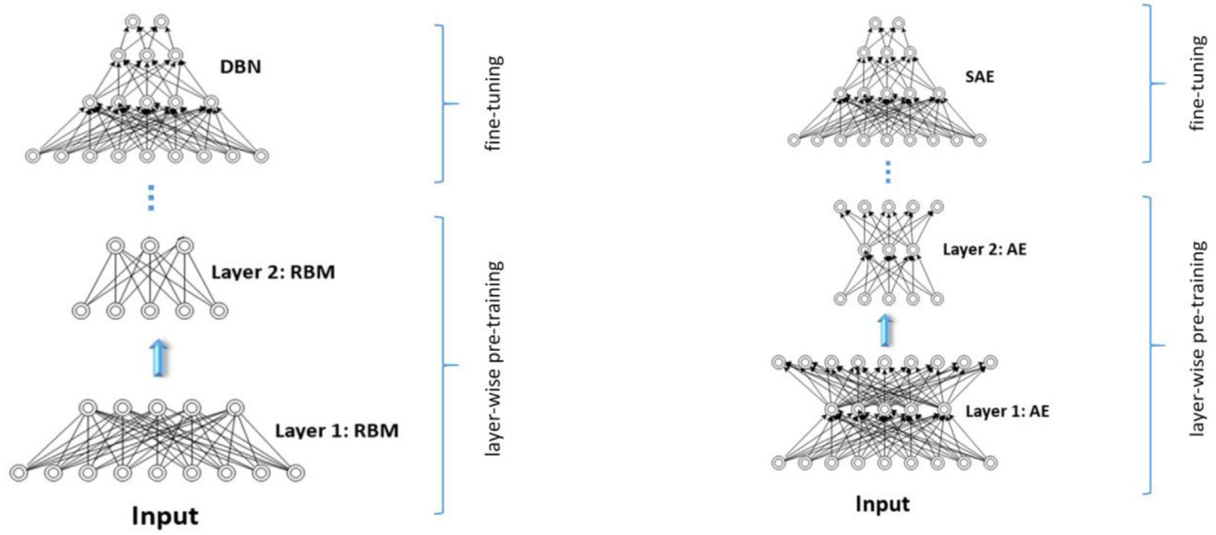


Fig. 1. Generic architecture of Convolutional Neural Network.



(a) Deep belief network (DBN)

(b) Stacked auto-encoder (SAE)

Fig. 2. A typical architecture showing layer-wise pre-training and fine-tuning procedures of (a) Deep belief network (DBN); (b) Stacked auto-encoder (SAE).

known as the encoding units. Fig. 2 (b) shows the architecture of an SAE as it gradually learns lower dimensional encoding units at each layer. A greedy layer-wise training algorithm is used to train any of DBN, DBM, or SAE networks, where the parameters of each layer are trained individually by keeping parameters in other layers fixed. After layer-wise training of all layers, also known as pre-training, the hidden layers are stacked together. The entire network with all the stacked layers is then fine-tuned against the target output units to adjust all the parameters for a classification task as illustrated in Fig. 2. DBNs and SAEs have achieved state-of-the-art performance in various vision-related applications such as face verification [45], phone recognition [46], and emotion recognition from image and speech [47,48]. Moreover, several studies [45,49] have combined the advantages of different deep learning models to further boost performance in these recognition tasks. For example, Lee et al. [49] have shown that combining convolution and weight sharing features of CNNs with the generative architecture of DBNs offers better classification performance on benchmark datasets such as MNIST and Caltech 101 [49]. The hybrid of CNN and DBN models, also known as the CDBN model, enables scaling to problems with large images without requiring an increase in the number of network parameters.

2.3. Variational autoencoders

Variational autoencoder (VAE) is a generative model that is designed to learn meaningful latent representations of input data. The VAE architecture is analogous to an autoencoder, where the deterministic hidden layer is replaced with a parameterizable distribution formulated by variational Bayesian inference. VAE is, therefore, represented by a directed graphical model consisting of an input layer, a probabilistic hidden layer, and an output layer to generate examples that are probabilistically similar to the input class. Kullback Leibler (KL) divergence is used as a constraint between the prior and posterior distribution to achieve a smooth transition in the hidden distributions between different classes. Variational Bayesian inference is used to construct a cost function for the neural network that connects the input and hidden layers before the output layer [36]. The parameterization of hidden layers for several classes can be represented as parameter vectors. Linear combinations of these class-specific vectors can be obtained to represent different types of input into a new output example. VAE has successful applications in image generation [50], motion prediction [51], text generation [52], and expressive speech generation [53].

2.4. Generative adversarial networks

Generative adversarial network (GAN) is another generative model that is capable of creating realistic data (typically images) from a given class. A GAN is composed of two competing networks: the generator and the discriminator. The generator aims to generate synthetic images from raw noise input that are as good as real images. The discriminator network has a binary target corresponding to ‘fake’ or ‘real’ inputs as it classifies real images against the synthetically generated ones. The entire pipeline of two networks is trained with two alternating goals. One goal is to update the discriminator to improve its classification performance while keeping the generator parameters fixed. The discriminator network yields low cost values when correctly classifying the generator examples as ‘fake’ against ‘real’ images. The other goal is to update the generator network by holding fixed parameters for the discriminator. Low cost values for the generator indicate that generated synthetic images are too real for the discriminator to classify as ‘fake’ [35]. Thus, the two networks compete against each other until an optimal point has been reached, which ensures that the fake examples are indistinguishable from real examples. As a generative network, GAN has applications similar to VAE, including image generation [54] and super resolution [55].

The GAN model does not have control over generating different variants of data. Conditional GAN (CGAN) model alleviates this shortcoming by adding the ground truth label as a parameter to the generator. This modification in CGAN allows the GAN model to generate new images from different classes. The CGAN discriminator also has an additional input and only returns ‘real’ when the input looks real and matches the corresponding input class provided in the generator [56]. The authors in [57] have extended the conditional GAN architecture to construct images from semantic label maps. Bidirectional GAN learns to simultaneously generate new images and estimate the latent parameters of existing images [58]. For a given input example, the hidden representation can be extracted. Then the underlying representation can be used to generate a new image of similar semantic quality. The BigBiGan architecture [59] is an improved bidirectional GAN that achieves state-of-the-art results in learning new image representation and also in image generation tasks.

Despite the popularity and success of GANs, they are frequently plagued by instability in training [60] and subject to underfitting and overfitting [61]. Several studies aimed at improving training stability and performance of GAN. The authors in [62] approach these problems with a weight normalization that they call spectral normalization. Wasserstein GAN (WGAN) is another modification that improves the training of GAN for generating more realistic new example images. The authors in [63] motivate the improvement of GAN with significant theoretical underpinning. The main difference between GAN and WGAN is that instead of providing a binary decision about generated images being ‘fake’ or ‘real’, the discriminator network evaluates the generated images using a continuous quality score between ‘fake’ and ‘real’. In [64], the authors consider weight clipping as a part of WGAN training. Weight clipping is considered a penalty on the norm of critic gradient, which has shown to improve training stability and image generation quality. In addition to WGAN, there are other studies that attempt to improve GAN. For example, least squares generative adversarial networks have improved network stability and performance [65]. They replace the standard GAN cross-entropy loss with least squares loss to resolve the vanishing gradient problem. Recently, vector quantization is applied to VAE to generate synthetic images of quality rivaling GAN while avoiding the aforementioned problems in training GAN [66].

2.5. Flow-based models

Flow models construct a decoder that is the exact inverse of the encoder module. This allows exact sampling from the inferred data distribution. In VAE, a distribution parameter vector is extracted by the encoder to define a new distribution that is sampled and decoded to generate an image. In a flow model, given a latent variable, the encoder defines a deterministic transformation into an output image. An early flow model, known as Nonlinear Independent Components Estimation (NICE) [67], is used to generate images with corrections to corrupt image regions, which is known as inpainting. The authors in [37] have extended NICE with several more complex invertible operations, including various types of sampling and masked convolution to perform image generation. Their proposed model is similar to conditional GAN as it can include an additional target class parameter to constrain the output image class. Another generative model called ‘GLOW’ uses generative flow with invertible convolutions [68] and is capable of generating realistic high-resolution human face images.

2.6. Generative models for speech

Several related generative models are applied in realistic speech synthesis. WaveNet [69] is an audio generation network based on deep autoregressive models that are used for image generation (e.g. PixelRNN [70]). WaveNet has no recurrent connections, which increases training speed at the cost of increasing the depth of the neural network. In WaveNet, a technique called dilated convolution has been found effective in exponentially increasing the context region with the depth of a neural network. WaveNet also utilizes residual connections as described in Section 3.1. Authors in [69] have used conditioning on WaveNet to enable text-to-speech (TTS) generation that yields the state-of-the-art performance when graded by human listeners. Waveglow [71] is another model that combines WaveNet and GLOW for frequency representation of input text sequences to generate realistic speech. Another model, known as the Speech Enhancement Generative Adversarial Network (SEGAN) [72], uses deep learning and avoids preprocessing of speech using spectral domain techniques. The authors use a convolutional autoencoder model enhanced input speech signal by training in a generative adversarial setting. Another work [73] modifies the SEGAN autoencoder model in the context of Wasserstein GAN to perform noise-robust speech enhancement.

2.7. Recurrent neural networks

Another variant of neural networks, known as the recurrent neural network (RNN), captures useful temporal patterns in sequential data such as speech to augment recognition performance. An RNN architecture includes hidden layers that retain the memory of past elements of an input sequence. Despite effectiveness in modeling sequential data, RNNs have challenges using the traditional backpropagation technique for training with a sequence of data with larger degrees of separation [38]. The long short-term memory (LSTM) networks alleviate this shortcoming with special hidden units known as “gates” that can effectively control the scale of information to remember or forget in the backpropagation [38]. Bidirectional RNNs [74] consider context from the past as well as the future to process sequential data to improve performance. This, however, can hinder real-time operation as the entire sequence must be available for processing. A modification to LSTM, called Gated Recurrent Unit (GRU) [75], has been introduced in the context of machine translation. The GRU has been shown to perform well on translation problems with short sentences. Several variations of LSTM including GRU are compared in [76]. The authors in [76] demonstrate experimentally that, in general, the original LSTM structure is superior for various recognition tasks. LSTM is a powerful model, however, recent advances in attention-based modeling have shown to have better performance than RNN models for sequential and context based information processing [39].

2.8. Attention in neural networks

The process of attention is an important property of human perception that greatly improves the efficacy of biological vision. The 'attention process' allows humans to selectively focus on particular sections of the visual space to obtain relevant information, avoiding the need to process the entire scene at once. Consequently, attention provides several advantages in vision processing [77]. The attention model drastically reduces computational complexity by limiting the processing space to the region of importance. Additionally, the performance of vision applications is improved as the attention model learns to identify regions of importance. Attention models also reduce noise by excluding irrelevant parts of the visual scene from processing. This selective fixation allows a contextual representation of the scene without 'clutter'. Hence, the application of such attention based neural network models is promising for speech and vision processing.

Early studies have introduced attention by means of saliency maps (e.g., for mapping of points that may contain important information in an image). A more recent attempt has introduced attention to deep learning models. A seminal study by Larochelle et al. [78] models attention in a third-order Boltzmann machine that is able to accumulate information of an overall shape in an image over several fixations. The model is only able to see a small area of an input image. Thus, it learns by gathering information through a sequence of fixations over parts of the image. To learn the sequence of fixations and the overall classification task, the authors in [78] have introduced a hybrid-cost for the Boltzmann machine. This model shows similar performance to deep learning variants that use the whole input image for classification. Another study [79] proposes a two-step system for an attention-based model. First, the whole input image is aggressively downsampled and processed to identify candidate locations that may contain important information. Next, each location is visited by the model in its original resolution. The information collected at each location is aggregated to make the final decision. Similarly, Denil et al. [80] have proposed a two-pathway model for object tracking, where one focuses on object recognition and the other pathway regulates the attention process.

However, learning 'where and when' to attend is difficult as it is highly dependent on the input and the task. It is also ill-defined in the sense that a particular sequence of fixations cannot be explicitly dictated as ground truth. Due to this challenge, most recent studies on deep learning with attention have employed reinforcement learning (RL) for regulating the attention aspect of the model. Accordingly, a seminal study by Mnih et al. [77] builds a RL policy on a two-path recurrent deep learning model to simultaneously learn the attention process and the recognition task. Based on similar principles, Gregor et al. [81] propose a recurrent architecture for image generation. The proposed architecture uses a selective attention process to trace outlines and generate digits similar to a human. Another study [82] utilizes the selective attention process for image captioning. In this study, the RL based attention process learns the sequence of glimpses through the input image that best describes the scene representation. Conversely, Mansimov et al. [83] leverage the RL based selective attention on an image caption to generate new images described in the caption. In this approach, the attention mechanism learns to focus on each word in a sequential manner that is most relevant for image generation. Despite impressive performance in learning selective attention using RL, deep RL still involves additional burdens in developing suitable policy functions that are extremely task-specific, and hence, are not generalizable. RL with deep learning also frequently suffers from instability in traini.

A different set of studies on neural networks analogous to the Turing machine architecture suggest the use of an attention process for interacting with external memory of the overall system. In this approach, the process of attention is implemented using a neural controller and a memory matrix [84]. The attentional focus allows selective access to the memory, which is necessary for memory control [84]. The neural Turing machine work is further explored in [85] considering attention-based global and local foci on an input sequence for machine translation. In [86], an attention mechanism is combined with a bidirectional LSTM network for speech recognition. In [87], the authors, inspired by LSTM for natural language processing (NLP), add a trust gate to augment LSTM in the application of human skeleton-based action recognition. Vaswani et al. [39] use an attention module called 'Transformer' to completely replace recurrency in language translation problems. This model is able to achieve improved performance on English-to-German and English-to-French translation. Zhang et al. [88] propose self-attention generative adversarial networks (SAGAN) for image generation. A standard convolutional layer can only capture local dependencies in a fixed shape window. Attention mechanism allows the discriminator and generators of the GAN model to operate over larger and arbitrarily shaped context regions [88].

2.9. Neural architecture search

Neural architecture search (NAS) involves automated selection of the architectural parameters of a neural network. In [89], architectural parameters including CNN filter size, stride, and the number of filters in a given convolutional layer are selected using NAS. Additionally, skip connections (discussed in Section 3.1) are automatically selected to generate densely connected CNN. The method in [89] uses RL to train an RNN to generate architectural parameters of a CNN. A more recent method, called the Differential Architecture Search (DARTS) [90], avoids the RL paradigm and formulates the problem of parameter selection as a differentiable function that is amenable to gradient descent. The gradient descent formulation improves the performance over RL and drastically reduces computational time for the search. Another work, known as the progressive neural architecture search [91], performs a search over CNN architectures. They begin with a simple structure and progress through a parameter search space toward more complex CNN models. They are also able to reduce the search time and space for the optimal architecture when compared to RL methods. They have reported the state-of-the-art performance on the CIFAR-10 image classification data set. In order to show the growth in deep learning models, Fig. 3 summarizes the literature search results with model names found in the article abstracts as of 2019. Section 3 elaborates on the contributions of these deep learning models to various speech and vision related applications.

3. Deep learning in speech and vision processing

This section discusses the impact of neural networks that are driving the state-of-the-art intelligent speech and vision systems.

3.1. Deep learning in computer vision

3.1.1. Image classification and scene labeling

The CNN model is first introduced to perform recognition of ten hand-written digits using image examples from the MNIST data

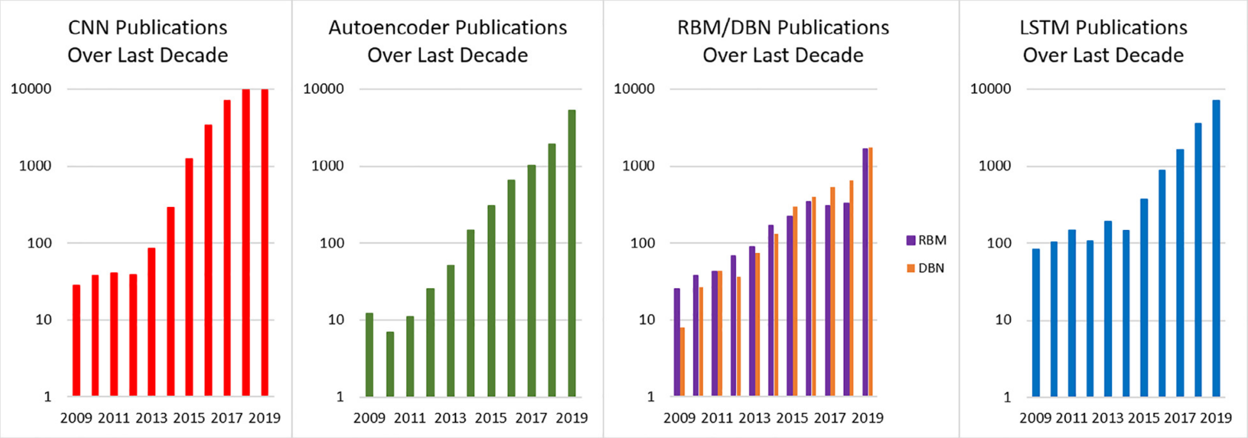


Fig. 3. Literature search showing increasing prominence of deep learning techniques.

set. The proposed CNN model has shown significant performance improvement in hand-written digit recognition task compared to earlier state-of-the-art machine learning techniques. Since then CNNs have seen several evolutions and the current versions of CNN are tremendously successful in solving more complex and challenging image recognition tasks [21,33,92,93]. For example, Krizhevsky et al. [33] utilize a deep CNN architecture named 'AlexNet' for solving the ImageNet classification challenge [94] to classify 1000 objects from high-resolution natural images. Their proposed CNN architecture has considerably outperformed previous state-of-the-art methods during the earliest attempt with the ImageNet classification challenge. The image recognition performance gradually improved as reported in several publications such as GoogleNet [93], VGGNet [95], ZFNet [96], and ResNet [97], following the initial success of AlexNet. More recently, He et al. [98] have extended AlexNet to demonstrate that a carefully trained deep CNN model is able to surpass human-level recognition performance, reported in [94] on the ImageNet data set. AlexNet [33] and GoogLeNet [93] are two of the pioneering CNN architectures that have significantly improved image classification performance compared to the conventional hand-engineered computer vision models. However, a limitation of these models is the vanishing gradient problem when increasing the number of layers to achieve more depth in learning abstract features. Consequently, a more sophisticated CNN architecture, such as ResNet [97], has been proposed by incorporating "residual block" in the architecture. A residual block combines a convolutional operation and a skip connection into an output. The skip connection directly passes the input with no transformation. This allows the model to achieve very deep structures providing a remedy to the vanishing gradient problem. Densely connected networks are introduced by Huang et al. [99]. They allow forward connections between any two convolutional layers called 'skip' connections. These connections between further-separated layers reduce vanishing gradient and improve the efficiency with reuse of features. Another architecture called squeeze-and-excitation [100] considers channel-wise dependencies in convolutional feature maps. This is performed by calculating and using the mean value for each channel to inform a rescaling of the feature maps. Recently, a technique called EfficientNet [101] is used for scaling of the CNN model. The authors first apply Neural Architecture Search (described in Section 2.8), and then uniformly scale network depth, width, and resolution simultaneously. This method has yielded the state-of-the-art performance in image recognition with an order of magnitude less parameters. The reduction in parameters here also implies faster inference. In Section 4, we extend this discussion of efficient networks for applications in limited resource environments. Scene labeling is another computer vision application that assigns target classes to multiple portions of an image based on the local content. Farabet et al. [92] have proposed a scene labeling method using a multi-scale CNN that yields record accuracies on several scene labeling problem datasets with up to 170 classes. CNNs have also demonstrated the state-of-the-art performance in other computer vision applications, such as in human face, action, expression, and pose recognitions. Table 1 shows performance error rates of the neural networks described above for image classification.

3.1.2. Human face, action, and pose recognition

Human-centric recognitions have long been an active area of research in computer vision. A recent approach in human face recognition is dedicated to improving the cost function of neural networks. The objective of such cost function for face recognition is to maximize interclass variation (facial variations between human individuals) and minimize intraclass variation (facial variations within an individual due to facial expressions). Wang et al. [102] have constructed a cost function called large margin cosine loss (LMCL), which achieves the desired variational properties. Using LMCL, their proposed model is able to achieve the state-of-the-art performance on several face recognition benchmarks. Following this work, Deng et al. [103] reformulate the cost function for face recognition. Their cost function, Additive Angular Margin Loss (ArcFace), is shown to further increase the margin between different face classes. Thus, ArcFace improves face recognition performance on a large experimental study of 10 datasets. Several CNN-based models are proposed in the literature to perform human action recognition. An architectural feature called temporal

Table 1
Summary of the significant state-of-the-art CNN image classification results (*actual class error within top 5 predictions, **pixel class error).

Architecture	Data set	Error rate
AlexNet [33] - University of Toronto 2012	ImageNet (natural images)	17.0%*
GoogLeNet [93] - Google 2014	ImageNet (natural images)	6.67%*
ResNet [97] - Microsoft 2015	ImageNet (natural images)	4.70%*
Squeeze & Excitation [100]– Oxford 2018	ImageNet (natural images)	2.25%*
Multiscale CNN [92] - Farabet et al. 2013	SIFT/Barcelona (scene labeling)	32.20%**

pyramid pooling is used in [104] to capture details from every frame in a video and is shown to perform action classification well with a small training set. Another architecture, called the two-stream CNN, analyzes both spatial and temporal contexts independently and gives competitive results on standard video action benchmarks [95]. CNN architectures that find pose features in an intermediate layer have been used for human action recognition. One of the more successful architectures for action recognition is called R*CNN [105]. This model uses contexts from the scene with human figure data to recognize actions. Action recognition has been performed using a skeletal representation of human individuals instead of RGB video of the entire body posture. Kinect [106] has been widely used to structure illumination of an individual to obtain a 3D skeleton measurement. Kinect skeletons are mapped to color images to represent 3D data and used as input in [107] for a ResNet CNN. Tang et al. [108] apply RL for a graph based CNN (GCNN) that captures structural and temporal information from 3D skeleton input. The authors note that future work may exploit the graph structure in the weight initialization process. Another approach [109] uses raw depth maps and intermediate 3D skeleton features in a multiple channel CNN. A fusion method is applied to the output of different CNN channels to leverage both modalities. This work improves accuracy on a benchmark with a large number of action classes.

CNNs are used in human pose estimation, for example, Deep Pose [20] is the first CNN application to perform pose estimation, which has outperformed earlier methods [110]. Deep Pose is a cascaded CNN based pose estimation framework. The cascading allows the model to learn an initial pose estimation based on the full image. A CNN based regressor is then used to refine the joint predictions using higher resolution sub-images. Tomson et al. [21] propose a 'Spatial Model' which incorporates a CNN architecture with Markov Random Field (MRF) and offers improved results in human pose estimation. Adversarial learning is applied to 2D images in [111] to extend the output pose prediction into 3D space. Furthermore, new sensing techniques allow efficient processing of 3D volumetric data using 3D CNNs. For example, in [112], human hand joint locations are estimated in real-time using a volumetric representation of input data and a 3D CNN. Another work extends pose estimation to dense pose estimation [113], where the goal is to generate a 3D mesh surface of an individual from 2D images.

3.1.3. Saliency detection and tracking

Saliency detection aims to identify regions of an input image that represent an object class of interest. Recent work in saliency detection has proposed the integration of a CNN with an RNN. For example, in [114], RNN is used to refine CNN feature maps by iteratively leveraging deeper contextual information than pure CNN. The work in [115] extends the idea of RNN feature map refinement by introducing multi-path recurrency, which is a feedback connection from different depths of CNN. Deep learning has also been applied to detect salient objects in video. One of the recent studies has used 3D CNN [116] to capture the temporal information, while another study [117] incorporates LSTM on top of CNN to capture the temporal information. Recently, Siamese CNN has been proposed to track objects over time in video frames. A Siamese CNN is a two branch CNN that takes two input images. The branches merge in the deep layers to produce an inference on the relation between the two images. In [118,119], Siamese CNN is used to generate information between adjacent image patches, which are used to track objects. RL is another technique that is applied in [120] for tracking biological image features at the subpixel level.

3.1.4. Image generation and inpainting

Generative models including VAE, GAN and its variants, and flow-based models have applications in image generation and image modification. As mentioned in Section 2.3–2.5, these generative models perform image generation and inpainting including human face image generation. These models are capable of several other applications. In [121], a method called cycle GAN is used for the unpaired image-to-image translation problem. Image-to-image translation would typically involve training on scenes where the input and output domains are given. For example, pairs of pictures of day and night at a location could be a training set. Then given a new image of a location in the day, the network outputs a night image. What cycle GAN accomplishes is even more impressive. The training is done without image pairs. So, the day and night images used in training are not from the same locations. The network is then trained to convert day images into night images. Another important GAN application is photo inpainting. When a part of an image is removed or distorted, the network can make a guess of the missing part, for example, face inpainting [122] or natural image inpainting [123]. A recent study has considered partial convolution to perform inpainting with irregularly removed regions [124]. A related application of GAN is semantic image generation. Parts of an image have semantic labels and the goal is to generate an image matching the labels. The authors in [57] use conditional GAN to generate high-resolution realistic images from semantic maps. A video prediction model based on flow networks has success comparable to VAE in short-period prediction of future frames [125].

Table 2 summarizes variants of CNN with their contributions and limitations in computer vision applications. A common obser-

Table 2
Comparison of Convolutional Neural Network models.

Architecture	Application	Contribution	Limitations
He et al. [98] AlexNet variant	Image classification	First human-level image classification performance (including fine grained tasks e.g. 100 dog breeds differentiation). Used ReLu for generalization and training	Misclassification of image cases that require context
Farabet et al. [92] multiscale CNN	Scene labeling	Weight sharing at multiple scales to capture context without increasing number of trainable parameters. Global application of graphical model to get consistent labels over the image	Does not apply unsupervised pretraining
Wang et al. [104] Temporal pyramid pooling CNN	Action recognition	Temporal pooling for action classification in videos of arbitrary length reduces the chance of overlooking important frames in decision	Challenging similar actions often misclassified
Tomson et al. [21] Joint CNN / graphical model	Human pose estimation	Combining MRF with CNN constrains plausible joint configurations to impact CNN body part detection	This model works well for limited set of human poses, general space of human poses remains a challenge
Ge et al. [112] 3D CNN	Human hand pose estimation	Volumetric processing of human depth maps of human hands using 3D CNN. 3D reasoning improves occluded finger estimation	Inherently constrained model. Requires clean and presegmented hand regions for pose estimation. The acceptable range of hand joint motion is limited

vation of these studies is that the proposed CNN models can yield human-level performance for simpler tasks. In [98], the authors note that image classification performance decreases when context is required for image understanding. A similar challenge is observed in human action recognition tasks using visual data or images. Authors in [104] have reported that similar human actions are more challenging to classify using neural network algorithms. In [21], the model only works well for a constrained set of human poses. When the classification problems become very difficult such as an arbitrary view or context dependent tasks, the architectures of vision algorithms still have room to improve.

3.2. Deep learning in speech recognition

In addition to offering excellent performance in image recognition [21,33,92,93], deep learning models have also shown state-of-the-art performance in speech recognition [126–128]. A significant milestone is achieved in acoustic modeling with the aid of DBNs at multiple institutions [127]. Following the work in [28], DBNs are trained in layer-wise fashion followed by end-to-end fine-tuning for speech applications as shown in Fig. 2 above. The DBN architecture and training process have been extensively tested on several large-vocabulary speech recognition datasets including TIMIT, Bing-Voice-Search speech, Switchboard speech, Google Voice Input speech, YouTube speech, and the English-Broadcast-News speech data set. DBNs significantly outperform state-of-the-art methods in speech recognition when compared to highly tuned Gaussian mixture model (GMM)Hidden Markov Model (HMM). SAEs likewise are shown to outperform GMM-HMM on Cantonese and other speech recognition tasks [43].

RNN has succeeded in improving speech recognition performance because of its ability to learn sequential patterns as seen in speech, language, or time-series data. RNNs have challenges in using the traditional backpropagation technique for training such models. This technique has difficulties to process portions of a sequence with long term separation in memory [39]. The problem is addressed with the development of long short-term memory (LSTM) networks that use special hidden units known as “gates” to retain memory over longer portions of a sequence [40]. Sak et al. [129] have first studied the LSTM architecture in speech recognition over a large vocabulary set. Their double-layer deep LSTM is found to be superior to a baseline DBN model. LSTM has been successful in an end-to-end speech learning method, known as Deep-Speech-2 (DS2), for two largely different languages: English and Mandarin Chinese. Other speech recognition studies using an LSTM network have shown significant performance improvement compared to previous state-of-the-art DBN based models. Furthermore, Chien et al. [130] have performed an extensive experiment with various LSTM architectures for speech recognition and compared the performance with state-of-the-art models. The LSTM model is extended in Xiong et al. [131] to bidirectional LSTM. This BLSTM is stacked on top of convolutional layers to improve speech recognition performance. The inclusion of attention enables LSTM models to outperform purely recurrent architectures. An attention mechanism called Listen, Attend, and Spell (LAS) is used to encode, attend, and decode, respectively. This LAS module is used with LSTM to improve speech recognition performance [132]. Using a pretraining technique [133] with attention and LSTM model, speech recognition performance has been improved to a new state-of-the-art level. Another memory network based on RNN is proposed by Weston et al. [134] to recognize speech content. This memory network stores pieces of information to retrieve the answer related to the inquiry, which makes it unique and distinctive from standard RNNs and LSTMs. RNN-based models have reached far beyond speech recognition to support Natural Language Processing (NLP). NLP aims to interpret language and semantics from speech or text to perform a variety of intelligent tasks, such as responding to human speech, smart assistants (Siri, Alexa, and Cortana), analyzing sentiment to identify positive or negative attitude towards a situation, processing events or news, and language translation in both speech and texts. To summarize key results in speech recognition using DBNs, RNNs (including LSTMs), and attention models, Table 3 presents different architectures, datasets, and performance achieved in the state-of-the-art literature.

Although RNNs/LSTMs are standard in sentiment analysis, authors in [135] have proposed a novel nonlinear architecture of multiple LSTMs to capture sentiments from phrases that constitute different order of the words in natural language. Researchers from Google machine learning [136] have developed a machine-based language translation system that runs Google’s popular online translation service. Although this system has been able to reduce average error by 60% compared to the previous system, it suffers from a few limitations. A more efficient translator, neural machine translator (NMT) [136], takes an entire sentence as input at one time, instead of sentences in parts. This improves the context and semantic representation by the model over traditional methods. More recently, a hybrid approach combines sequential language patterns from LSTMs and hierarchical learning of images from CNNs. This hybrid approach can describe image content and context using natural language descriptions. Karpathy et al. [137] have introduced this hybrid approach for image captioning to incorporate both visual data and language descriptions to achieve optimal performance in image captioning across several datasets. Table 4 summarizes variants of RNNs and their contributions and limitations in state-of-the-art speech recognition systems.

Similar to vision applications, RNN models can yield human-level performance for simpler speech recognition tasks. For both CNNs and RNNs, the architecture is inherently driven by the problem domain. Examples of such applications include: 1) multiscale CNN to gather context for labeling across a scene [92], 2) temporal pooling to understand actions across time [104], 3) MRF graphical modeling on top of CNN to constrain plausible joint configurations [21], 4) long term memory components for context retrieval in stories, and 5) CNN fused with RNN to interpret images using language. In [134], the authors note that the question and input

Table 3

Summary of the significant state-of-the-art DNN speech recognition models (*perpeplexity-size of model needed for optimal next word prediction with 10 K classes, **word error rate).

Architecture	Data set	Error rate
RNN [126] - FIT, Czech Republic, Johns Hopkins University, 2011	Penn corpus (natural language modeling)	123*
Autoencoder/DBN [127] - Collaboration, 2012	English broadcast news speech corpora (spoken word recognition)	15.5%**
LSTM [129] - Google, 2014	Google voice search task (spoken word recognition)	10.7%**
Deep LSTM [130] - National Chiao Tung University, 2016	CHiME 3 challenge (spoken word recognition)	8.1%**
CNN-BLSTM [131] - Microsoft, 2017	Switchboard (spoken word recognition)	5.1%
Attention (LAS) & LSTM [132] - Google, 2018	In-house google dictation (spoken word recognition)	4.1%
Attention & LSTM with pretraining [133] - Collaboration, 2018	LibriSpeech (spoken word recognition)	3.54%

Table 4
Comparison of recurrent neural network models in speech processing.

Architecture	Application	Contribution	Limitations
Amodei et al. [159] Gated Recurrent Unit (GRU) network	English or Chinese speech recognition	Optimized speech recognition using GRUs to achieve near human-level results	Deployment requires GPU server
Weston et al. [134] Memory network	Answering questions about simple text stories	Integration of long term memory (readable and writable) component within neural network architecture	Questions and input stories are still rather simple
Wu et al. [136] Deep LSTM	Language translation (e.g. English-to-French)	Multi-layer LSTM with attention mechanism	Challenging translation cases and multi-sentence input yet to be tested
Karpathy et al. [137] CNN/RNN fusion	Labeling images and image regions	Hybrid CNN-RNN model to generate natural language descriptions of images	Fixed image size / requires training CNN and RNN models separately

stories are rather simple for the neural models to handle. In [136], the authors report that challenging translation problems are yet to be successfully addressed in current studies. As tasks become more complex or highly abstract, more sophisticated IS are required to reach human-level performance.

Speech emotion and visual speech recognition are two important topics that have gained recent attention in the deep learning literature. Mirsamadi et al. [138] have used a deep recurrent network with local attention to automatically learn speech features from audio signals. Their proposed RNN captures a large context region while the attention focuses on the aspects of speech relevant to emotion. This idea is later extended in Chen et al. [139], where operation on frequency bank representation of speech signals can be used as inputs into a convolutional layer. This convolutional layer is followed by LSTM and attention layers. Mirsamadi et al. have further improved the work of Chen et al. to yield the state-of-the-art performance on Interactive Emotional Dyadic Motion Capture (IEMOCAP) emotion recognition tasks. Another work in [140] applies adversarial autoencoder for emotion recognition in speech. However, they use heuristic features as network input including spectral and energy features of speech in the IEMOCAP emotion recognition task.

Visual speech recognition involves lip reading of human subjects in video data to generate text captions. Recently, two notable studies have used attention-based networks for this problem. Afouras et al. [141] use 3D CNN to capture spatio-temporal information of the face, and a transformer self-attention module for speech recognition from extracted convolutional features. Stafylakis et al. [142] consider zero-shot keyword spotting, where the phrase is not seen in training data and is searched in a visual speech video. The input video is first fed into a 3D spatial–temporal residual network to capture face information over time. This is followed by attention and LSTM layers that detect the phrase in the video. Both studies consider “in the wild” speech recognition or a large breadth of natural sentences in speech.

3.3. Datasets for speech and vision applications

Several current datasets have been compiled for state-of-the-art benchmarking of computer vision. ImageNet is a large-scale data set of annotated images including bounding boxes. This data set includes over 14 million labeled images spanning more than 20,000 categories [94]. CIFAR-10 is a data set of smaller images that contain a recognizable object class in low resolution. Each image is only 32x32 pixels, and there are 60,000 images for each of 10 classes [143]. Microsoft Common Objects in Context (COCO) provides segmentation of objects in images for benchmarking problems including saliency detection. This data set includes 2.5 million instances of objects in 328 thousand images [144]. More complex image datasets are now being developed for unmanned aerial vehicle (UAV) deployment. Here, detection and tracking take place in a highly unconstrained environment. This includes different weather, obstacles, occlusions, and varying camera orientations relative to the flight path. Recently, two large scale datasets are released for benchmarking detection and tracking in UAV applications. The Unmanned Aerial Vehicle Benchmark [145] includes single and multiple bounding boxes for detection and tracking in various flight conditions. An ambitious project called Vision Meets Drones [146] gathered a data set with 2.5 million object annotations for detection and tracking in UAV urban and suburban flight environments.

Speech recognition also has several current datasets for state-of-the-art benchmarking. Defense Advanced Research Projects Agency (DARPA) has commissioned a collaboration between Texas Instruments and MIT (TIMIT) to make a speech transcription data set. The TIMIT data set includes 630 speakers from several American English dialects [147]. VoxCeleb is a more current speech data set with 1000 celebrities’ voice transcriptions in a more unconstrained or “in the wild” setting [148]. In machine translation, Stanford’s NLP group has released several public language translation datasets including WMT’15 English-Czech, WMT’14 English-German, and IWSLT’15 English-Vietnamese. The English to Czech and English to German datasets have 15.8 and 4.5 million sentence pairs, respectively [149]. CHiME 5 [150] is a speech recognition data set that contains challenging speech recognition conditions including multiple speaker natural conversations. A data set called LRS3-TED has been compiled for visual speech recognition [151]. This data set includes hundreds of hours of TED talk videos with subtitles aligned in time at the resolution of single words. Many other niche datasets on computer speech and vision can be found on the Kaggle Challenge website free to the public.

3.4. Deep learning in commercial speech and vision applications

In recent years, giant companies such as Google, Facebook, Apple, Microsoft, IBM, and others have adopted deep learning as one of their core areas of research in artificial intelligence (AI). Google Brain [152] focuses on engineering the deep learning methods, such as tweaking CNN-based architectures to obtain competitive recognition performance in various challenging vision applications. They use a large number of cluster machines and high-end GPU-based computers to parallelize their computation. Facebook conducts extensive deep learning research in their Facebook AI Research (FAIR) [153] lab for image recognition and natural language understanding. Many users around the globe are already taking advantage of this recognition system in the Facebook application. Their next milestone is to integrate the deep learning-based NLP approaches to the Facebook application to achieve near human-level performance in understanding language. Recently, Facebook has launched a beta AI assistant system called ‘M’ [154]. ‘M’ utilizes NLP to support more complex tasks such as purchasing items, arranging delivery of gifts, booking restaurant reservations, and making travel arrangements or appointments. Microsoft has investigated Cognitive toolkit [155] to show efficient ways to train deep models across distributed computers. They have also implemented an automatic speech recognition system that achieves human-level conversational speech recognition [156]. More recently, they have introduced a deep learning-based speech invoked assistant called Cortana [157]. Baidu has studied deep learning to create massive GPU systems with Infiniband [158] networks. Their speech recognition system named Deep Speech 2 (DS2) [159] has shown remarkably improved performance over its competitors. Baidu is also one of the pioneering research groups to introduce deep learning-based self-driving cars with BMW. Nvidia has invested efforts in developing state-of-the-art GPUs to support more efficient and real-time implementation of complex deep learning models [160]. Their high-end GPUs have led to one of the most powerful end-to-end solutions for self-driving cars. IBM has recently introduced their cognitive system known as Watson [161]. This system incorporates computer speech and vision recognition in a human friendly interface and NLP backend. Traditional computer models have relied on rigid mathematic principles to utilize software built upon rules and logic. Instead, Watson relies on what IBM is calling "cognitive computing". The Watson-based cognitive computing system has already been proven useful across a range of different applications such as healthcare, marketing, sales, customer service, operations, HR, and finance. Other major tech companies that are actively involved in deep learning research include Apple [162], Amazon [163], Uber [164], and Intel [165]. Fig. 4 summarizes publication statistics over the past 10 years searching article abstracts for 'deep learning', 'computer vision', 'speech recognition', and 'natural language processing'.

Although deep learning has revolutionized today's IS with the aid of computational resources, its application in more personalized settings, such as embedded and mobile hardware systems, is another challenge that has led to an active area of research. This challenge is due to the extensive requirement of high-powered and dedicated hardware for executing the most robust and sophisticated deep learning algorithms. Consequently, there is a growing need for developing more efficient, yet robust deep models in resource restricted hardware environments. The next sections survey some recent advances in highly efficient deep models that are compatible with mobile hardware systems.

4. Speech and vision on resource restricted hardware platforms

The success of future speech and vision systems depends on accessibility and adaptability to a variety of platforms that eventually drive the prospect of commercialization. While some platforms are intended for public and personal use, there are other commercial, industrial, and online-based platforms - all of which require seamless integration of IS. However, state-of-the-art deep learning models have challenges in adapting to embedded hardware due to large memory footprint, high computational complexity, and high-power consumption. This has driven research on improving system performance of compact architectures in resource restricted platforms. The following sections highlight some of the major research efforts in integrating sophisticated algorithms in resource restricted user platforms.

4.1. Speech recognition on mobile platforms

Handheld devices such as smartphones and tablets are ubiquitous in modern life. Hence, a large effort in developing IS is dedicated to mobile platforms with a view to reaching out to billions of mobile users around the world. Speech recognition has been a pioneering application in developing smart mobile assistants. The voice input of a mobile user is first interpreted using a speech recognition algorithm. The answer is then retrieved by an online search. The retrieved information is then spoken out by the virtual mobile assistant. Major technology companies, such as Google [166], have enabled voice-based content search on Android devices and a similar voice-based virtual assistant, known as Siri, is also available with Apple's iOS devices. This intelligent application provides mobile users with a fast and convenient hands-free feature to retrieve information.

However, mobile devices, like other embedded systems, have computational limitations and issues related to power consumption and battery life. Therefore, mobile devices usually send input requests to a remote server to process and send the information back to the device. This further brings in issues related to latency due to wireless network quality while connecting to the server. As an example, Keyword spotting (KWS) [167] detects a set of previously defined keywords from speech data to enable hands-free features in mobile devices. The authors in [167] have proposed a low-latency keyword detection method for mobile users using a deep learning-based technique and termed it as 'deep KWS'. The deep KWS method has not only been proven suitable for low-powered embedded systems but also has outperformed the baseline Hidden Markov Models for both noisy and noise-free audio data. The deep KWS uses a fully connected DNN with transfer learning [167] based on speech recognition. The network is further optimized for KWS with end-to-end fine-tuning using stochastic gradient descent. Sainath et al. [168] have introduced a similarly small footprint KWS system based on CNNs. Their proposed CNN uses fewer parameters than a standard DNN model, which makes the proposed system more attractive for platforms with resource constraints. Chen et al. [169] in another study propose the use of LSTM for the KWS task. The inherent recurrent connections in LSTM can make the KWS task suitable for resource restricted plat-

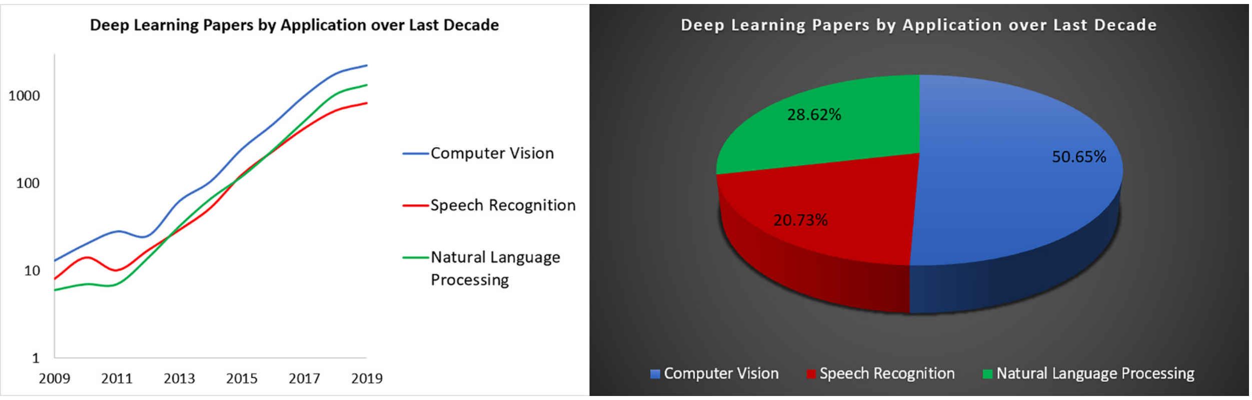


Fig. 4. Trends of deep learning applications in the literature over the last decade.

forms by improving computational efficiency. To support this, the authors further show that the proposed LSTM outperforms a typical DNN-based KWS method. A typical framework for deep learning based KWS system is shown in Fig. 5.

Similar to KWS systems, automatic speech recognition (ASR) [170] has become increasingly popular with mobile devices as it alleviates the need for tedious typing on small mobile devices. Google provides ASR-based search services [166] on Android, iOS, and Chrome platforms. Apple iOS devices are equipped with a conversational assistant named Siri. Mobile users can also type texts or emails by speech on both Android and iOS devices [171]. However, ASR service is contingent on the availability of cellular mobile network since the recognition task is performed on a remote server. This is a limitation since mobile network strength can be low, intermittent, or even absent at places. Therefore, developing an accurate speech recognition system in real-time, embedded on standalone modern mobile devices, is still an active area of research.

Consequently, embedded speech recognition systems using DNNs have gained attention. Lei et al. [170] have achieved substantial improvement in ASR performance over traditional GMM-based acoustic models even at a much lower footprint and memory requirements. The authors show that a DNN model, with 1.48 million parameters, outperforms the generic GMM-based model while exploiting only 17% of the memory used by GMM. Furthermore, the authors use a language model compression scheme LOUDS [172] to gain a further 60% improvement in the memory footprint for the proposed method. Wang et al. [173] propose another compressed DNN-based speech recognition system that is suitable for use in resource restricted platforms. The authors train a standard fully connected DNN model for speech recognition, compress the network using a singular value decomposition method, and then use split vector quantization algorithms to enhance computational efficiency. The authors have achieved a 75% to 80% reduction in memory footprint to a mere 3.2 MB. Additionally, they achieved a 10% to 50% reduction in computational cost with performance comparable to that of the uncompressed version. In [174], the authors show that low-rank representation of weight matrices can increase representational power per number of parameters. They also combine this low-rank technique with ensembles of DNN to improve the performance of KWS task. Table 5 summarizes small footprint speech recognition and KWS systems, which are promising for application in resource restricted platforms.

4.2. Computer vision on mobile platforms

Real-time recognition of objects or humans is a highly desirable feature with handheld devices for convenient authentication, identification, navigational assistance. When combined with speech recognition, it can even be used as a mobile teaching assistant. Though deep learning has advanced in speech recognition tasks on mobile platforms, image recognition systems are still challenging to deploy in mobile platforms due to resource constraints.

In a study, Sarkar et al. [175] use a deep CNN for face recognition in mobile platforms for user authentication. The authors first identify disparities in hardware and software between mobile devices and typical workstations in the context of deep learning, such as the unavailability of powerful GPUs and CUDA (an application programming interface by Nvidia that enables general-purpose processing in GPU) capabilities. The study subsequently proposes a pipeline that leverages AlexNet [33] through transfer learning [176] for feature extraction and then uses a pool of SVMs for scale-invariant classification. The algorithm is evaluated and compared in terms of runtime and face recognition accuracy on several mobile platforms embedded with Qualcomm Snapdragon CPUs and Adreno GPUs. The algorithm has achieved 96% and 88% accuracies on two standard datasets, UMD-AA [177] and MOBIO [178], respectively, with a minimum runtime of 5.7 s on the Nexus 6 mobile phone. In another study, Howard et al. [179] have introduced a class of efficient CNN models termed ‘MobileNets’ for mobile and embedded vision processing applications. MobileNet models leverage the layerwise separability in convolution

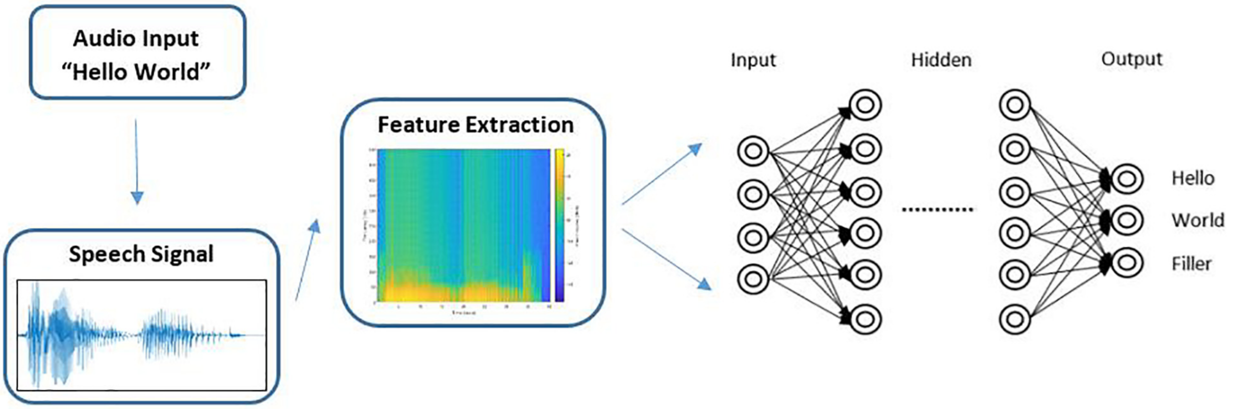


Fig. 5. Generalized framework of a keyword spotting (KWS) system that utilizes deep learning.

Table 5
KWS architectures with reduced computational and memory footprint (*relative improvement over comparison network from ROC curve, **WER (word error rate), ***relative FER (frame Error Rate) over comparison network).

Compression technique	Memory reduction	Error rate (varied datasets)
DNN improvement over HMM, 2014 [167]	2.1 M parameters	45.5% improvement*
CNN improvement over DNN, 2015 [168]	65.5 K parameters	41.1% improvement*
Fixed length vector LSTM, 2015 [169]	152 K parameters	86% improvement*
Split vector quantization, 2015 [173]	59.1 MB to 3.2 MB	15.8%**
Low rank matrices / ensemble training, 2016 [174]	400 nodes per layer to 100 nodes per layer	−0.174***

operation to obtain substantial improvements in efficiency over conventional CNNs. The study also defines two global hyperparameters that configure the width and depth of the MobileNet architecture, compromising between latency and accuracy in the model performance. The authors show approximately seven-fold reduction in trainable parameters using MobileNet at the cost of losing only 1% accuracy in multiple vision tasks when compared to conventional architectures. Su et al. [180] have further improved MobileNet by reducing model-level and data-level redundancies that exist in the architecture. Specifically, the authors suggest an iterative pruning strategy [181] and a quantization strategy [182] to address model-level and data-level redundancy, respectively. The authors show comparable accuracy of the proposed model using a conventional AlexNet on an ImageNet classification task with just 4% use of trainable parameters and 31% of computational cost per image inference.

Lane et al. [183] have also performed an initial study using two popular deep learning models: CNN and fully connected deep feedforward networks. These models are used to analyze audio and image data on three hardware platforms: Qualcomm Snapdragon 800, Intel Edison, and Nvidia Tegra K1 as these are commonly used in wearable and mobile devices. The study includes extensive analyses on energy consumption, processing time, and memory footprint on several state-of-the-art models such as Deep KWS, DeepEar, ImageNet [33], and SVHN [184] (street-view house number recognition) for speech and image recognition applications. The study identifies a critical need for optimization of these sophisticated deep models in terms of computational complexity and memory usage for effective deployment in regular mobile platforms.

In another study, Lane et al. [185] discuss the feasibility of incorporating deep learning algorithms in mobile sensing for a number of signal and image processing applications. They highlight the limitation that deep models for mobile applications are still implemented on cloud-based systems rather than on standalone mobile devices due to large computational overhead. However, the authors point out that mobile architectures have been advancing in recent years and may soon be able to accommodate complex deep learning methods in devices. The authors subsequently implement a DNN architecture on the Hexagon DSP of a Qualcomm Snapdragon SoC (standard CPU used in mobile phones). They compare its performance with classical machine learning algorithms such as decision tree, SVM, and GMM in processing activity recognition, emotion recognition, and speaker identification. They report increased robustness in performance with acceptable levels of resource use for the proposed DNN implementation in mobile hardware.

4.3. Compact, efficient, low power deep learning for lightweight speech and vision processing

As discussed in Sections 4.1 and 4.2, hardware constraints pose a major challenge in deploying the most robust deep models in mobile hardware platforms. This has led to a recent research trend that aims to develop compressed but efficient versions of deep models for speech and vision processing. One seminal work in this area is the development of the software platform ‘DeepX’ by Lane et al. [186]. ‘DeepX’ is based on two resource control algorithms. First, it decomposes large deep architectures into smaller blocks of sub-architectures and then assigns each block to the most efficient local processing unit (CPUs, GPUs, LPUs). Furthermore, the proposed software platform is capable of dynamic decomposition and resource allocation using a resource prediction model [186]. Deploying on two popular mobile platforms, Qualcomm Snapdragon 800 and Nvidia Tegra K1, the authors report impressive improvements in resource use by DeepX for four state-of-the-art deep architectures: AlexNet [33], SpeakerID [187], SVHN [188], and AudioScene in object, face, character, and speaker recognition tasks, respectively [186].

Sindhwani et al. [189], on the other hand, propose a memory efficient method using a mathematical framework to represent large dense matrices such as neural network parameters (weight matrices). Structured matrices, such as Toeplitz, Vandermonde, Couchy [190], utilize various parameter sharing mechanisms to represent a $m \times n$ matrix with much less than mn parameters [189]. Authors also show that the use of structured matrices results in substantial improvements in computations, especially in the matrix multiplication operations encountered in deep architectures. The computation time complexity $O(mn)$ is reduced to $O(mlog(n))$ [189]. This makes both forward computations and backpropagation faster and efficient while training neural networks. The authors test the proposed framework on a deep KWS architecture for mobile speech recognition and compare with other similar small footprint KWS models [168]. The results show that Toeplitz based compression gives the best model computation time, which is 80 times faster than the baseline at the cost of only 0.4% performance degradation. They also conclude that the compressed model has achieved a 3.6 times reduction in memory footprint compared to the small footprint model proposed in [168].

Han et al. [181] propose a neural network-based three-stage compression scheme known as ‘deep compression’ for reduction of memory footprint. The first stage called pruning [191] removes weak connections in a DNN to obtain a sparse network. The second stage involves trained quantization and weight sharing applied to the pruned network. The third stage uses Huffman coding for lossless data compression in the network. Authors report reduced energy consumption and a significant computing speedup in a comparison between various workstations and mobile hardware platforms. An architecture called ShuffleNet [192] uses two architectural features. In this architecture, group convolution, introduced in [33], is used with channel shuffle architecture in a novel way to improve the efficiency of convolutional networks. The group convolution improves the speed in processing images and offers comparable performance with reduced model complexity. Table 6 summarizes results from different studies for compressed network energy consumption executing AlexNet on a Tegra GPU. Fig. 6 on the next page summarizes publication statistics over the past five years on small footprint analysis of deep learning methods for ‘computer vision’, ‘speech processing’, and ‘natural language processing’ in resource restricted hardware platforms.

Table 6
Compressed architecture energy and power running AlexNet on a Tegra GPU.

Compression technique	Execution time	Energy consumption	Implied power consumption
Benchmark study, 2015 [185]	49.1msec	232.2 mJ	4.7 W (all layers)
Deep X software accelerator, 2016 [186]	866.7msec (average of 3 trials)	234.1 mJ	2.7 W (all layers)
DNN, various techniques, 2016 [181]	4003.8msec	5.0 mJ	0.0012 W (one layer)

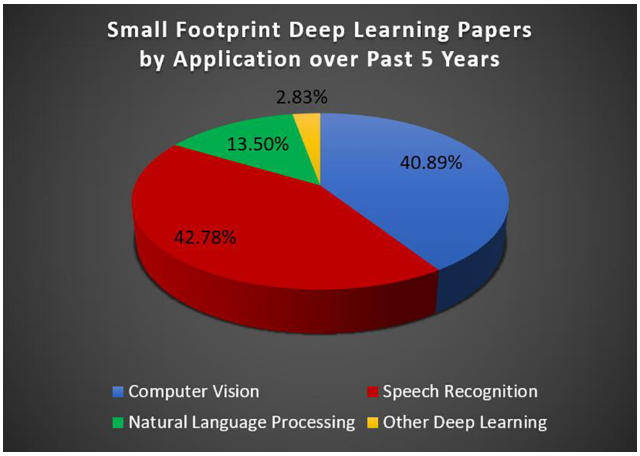


Fig. 6. Publications on small footprint implementations of deep learning in computer and speech and vision processing.

4.4. High-end hardware for neural network applications on mobile platforms

As architectures become more efficient, hardware on mobile devices is becoming more powerful and tailored to neural network applications. Qualcomm Snapdragon 865 is a high-end smartphone and tablet mobile processor. The processor incorporates one prime core and three additional fast ARM performance cores. Snapdragon 865 has integrated GPU that allows the processor to provide superior performance in GPU intensive graphics tasks [193]. Qualcomm considers Snapdragon 865 the next generation intelligent mobile platform. The company provides an on-device AI engine that improves the performance of the camera, battery life, audio, security, and gaming. Furthermore, Qualcomm provides AI software packages such as Neural Processing (NP) SDK, the Hexagon NN, common NN framework support, and Android NN API for deploying AI models in the device [194]. Apple's new A13 bionic chip includes a 64-bit ARM-based system which outperforms Qualcomm's latest Snapdragon 865 mobile processor and other high-end mobile processors such as Exynos 990 and MediTek Dimensity 1000 [195,196]. Snapdragon 865 includes Adreno 640 GPU with a performance similar to that of the mobile GPU in the A13 chip. Note that Apple has developed the Metal 2 software to optimize graphics and gaming on A13. More importantly, the software supports general purpose GPU (GPGPU) computing across this platform. Apple's Core ML framework [197] now supports custom machine learning models to run on iOS devices. This framework can accelerate these models up to nine times faster using just a tenth of the energy compared to running on regular GPU and CPU platforms [198]. Consequently, Apple's A13 bionic chip is the leading mobile chip in its category for the current market.

4.5. Edge computing for mobile platforms and IoT

The advancement of hardware platforms on mobile device now allows vendors to decentralize data processing off the cloud. Edge computing refers to the migration of computing closer to the edge of the network that includes the acquisition and consumption points [199,200]. The surge in edge computing is fueling the recent growth in internet of things (IoT) devices and applications that require real-time computing and faster communication channels. Consequently, IoT offers several benefits through the implementation of neural network architectures for speech and vision using edge computing. These benefits include improved run-time due to reduction of communication overhead [201,202], energy efficiency by proper management of computational resources [203,204], and improved memory efficiency with cloud offloading [205]. The performance of edge computing largely relies on the improved data processing capability of edge devices. This has created a demand for high performance computing hardware that is capable of handling large scale deep learning models on edge devices. One such popular device is the Intel® neural compute stick [206] which is a USB plug-and-play development kit. This device contains an Intel Movidius™ Myriad™ X Vision Processing Unit (VPU) and supports most of the popular deep learning frameworks. This low power USB device facilitates the development and deployment of CNN for vision applications with real-time performance. Nvidia Jetson [207] is another popular series of edge computing devices designed for deep learning. Nvidia Jetson offers a product line of low power stand-alone processing devices that are primed for deep learning applications supported by GPU-like parallelized processing.

The next section brings together the neural network architectures from Section 2, the computer speech and vision models from Section 3, and the mobile algorithms and hardware from Section 4 to discuss emerging applications of these systems.

5. Emerging applications of intelligent speech and vision Systems

We identify three fields of research that are shifting paradigm through recent advances in speech and vision-related frameworks. First, the quantification of human behavior and expressions from visual image and speech offers great potentials in cybernetics, security and surveillance, forensics, quantitative behavioral science, and psychology research [208]. Second, the field of transportation research is rapidly incorporating intelligent vision systems for smart traffic management and self-driving technology. Third, neural networks in medical image analysis show tremendous promise for 'precision medicine'. This represents a vast opportunity to automate clinical measurements, optimize patient outcome predictions, and assist physicians in clinical practice.

5.1. Intelligence in behavioral science

The field of behavioral science widely uses human annotations and qualitative screening protocols to study complex patterns in human behavior. These traditional methods are prone to error due to high variability in human rating and qualitative nature in behavioral information processing. Many computer vision studies on human behavior, e.g., facial expression analyses [209], can move across disciplines to revolutionize human behavioral studies with automation and precision.

In behavioral studies, facial expressions and speech are two of the most common means to detect emotional states of humans. Yang et al. use quantitative analysis of vocal idiosyncrasy for screening depression severity [23]. Children with neurodevelopmental disorders such as autism are known to have distinctive characteristics in speech and voice [24]. Hence, computational methods for detecting differential speech features and discriminative models [25] can help in the development of future applications to recognize emotion from the voice of children with autism. Recently, deep learning frameworks have been employed to recognize emotion from speech data promising more efficient and sophisticated applications in the future [26,27,210].

On the other hand, visual images from videos are used to recognize human behavioral contents [211] such as facial expressions, head motion, human pose, and gestures to support a variety of applications for security, surveillance, and forensics [212–214] and human–computer interactions [19]. The vision-based recognition of facial action units defined by facial action coding system (FACS) [215] has enabled more fine-grain analysis of emotional and physiological patterns beyond prototypical facial expressions such as happiness, fear, or anger. Several commercial applications for real-time and robust facial expression and action unit level analysis have recently appeared in the market with companies such as Noldus, Affectiva, and Emotient. With millions of facial images available for training, state-of-the-art deep learning models have enabled unprecedented accuracies in these commercially available facial expression recognition applications. These applications are designed to serve a wide range of research studies including classroom engagement analysis [216], consumer preference study in marketing [217], behavioral economics [218], atypical facial expression analysis in neurological disorders [219,220], and other work in the fields of behavioral science and psychology. The sophistication in face and facial expression analyses may unravel useful markers in diagnosing or differentiating individuals with behavioral or affective dysfunction such as those with autism spectrum disorder [221]. Intelligent systems for human sentiment and expression recognition will play lead roles in developing interactive human–computer systems and smart virtual assistants in the near future.

5.2. Intelligence in transportation

Intelligent transportation systems (ITS) cover a broad range of research interests including monitoring driver's inattention [1], providing video-based lane tracking and smart assistance to driving [2], monitoring traffic for surveillance and traffic flow management [3], and more recently developing self-driving cars [4]. Bojarski et al. have recently used deep learning frameworks such as CNN to obtain steering commands from raw images captured by a front-facing camera [5]. The system is designed to operate on highways, without lane markings, and in places with minimal visual guidance. Lane change detection [2,6] and pedestrian detection [7] have been studied in computer vision and are recently being added as safety features in personal vehicles. Similarly, computer vision assisted prediction of traffic characteristics, automatic parking, and congestion detection may significantly ease our efforts in traffic management and safety. Sophisticated deep learning methods, such as LSTM, are being used to predict short term traffic [6], and other deep learning frameworks are being used for predicting traffic speed and flow [8], and for predicting driving behavior [9]. In [10], the authors suggest several aspects of transportation that will be impacted by intelligent systems. Considering multimodal data collection from roadside sensors, RBM will be useful as this model is proven to handle multimodal data processing. Considering onboard vehicle systems, CNN can be combined with LSTM to take action in real-time to avoid accidents and improve vehicle efficiency. In line with these research efforts, several car manufacturing companies, such as Audi [222] and Tesla [223], are in active competition for developing next-generation self-driving vehicles with the aid of recent developments in neural network based deep learning techniques. Ride hailing and sharing is another growing domain in transportation. In ride hailing, there is a significant value in predicting pickup demand at different locations to optimize the transportation system and service. CNN has been recently used for location-specific demand of service prediction [224]. Travel time prediction has been performed using CNN and RNN to utilize road network topology and historical trip data [225]. Popular ride sharing services may benefit from recent advances in reinforcement learning. Alabbasi et al. have used deep Q-network (a model based on reinforcement learning) along with CNN to develop an optimal vehicle dispatch policy that ultimately improves traffic congestion and emission [226].

5.3. Intelligence in medicine

Despite tremendous development in medical imaging techniques, the field of medicine heavily depends on manual annotations and visual assessment of a patient's anatomy and physiology from medical images. Clinically trained human eyes sometimes miss important and subtle markers in medical images resulting in misdiagnosis. Misdiagnosis or even failure to diagnose early can lead to fatal consequences as misdiagnosis is known as the third most common cause of death in the United States [227]. The sophisticated deep learning models with the support of massive records of multi-institutional imaging databases may ultimately drive the future of precision medicine. Deep learning methods have been successful in medical image segmentation [11], shape and functional measurements of organs [14], disease diagnosis [12], biomarker detection [13], patients' survival prediction from images [228], and many more applications. Authors in [229] have used a hybrid of LSTM and CNN model to predict patient survival from echocardiographic videos of the heart motion, which has shown a prediction accuracy superior to that of trained cardiologists. Advances in deep neural networks have shown tremendous potential in almost all areas of medical imaging such as ophthalmology [230], dental radiography [231], skin cancer imaging [232], brain imaging [233], cardiac imaging [234,235], urology [236], lung imaging [237], stroke imaging [238], and so on. In addition to academic research, many commercial companies, such as Philips, Siemens, and IBM are investing on large initiatives towards incorporating deep learning methods in intelligent medical image analysis. However, a key challenge remaining is the requirement of large ground truth medical imaging data annotated by clinical experts. With commercial initiatives, clinical and multi-institutional collaborations, deep learning-based applications may soon be available in clinical practice.

6. Limitations of deep computational models

Despite unprecedented successes of neural networks in recent years, we identify a few specific areas that may greatly impact the future progress of deep learning in intelligent systems. The first area is to develop a robust learning algorithm for deep models that requires a minimal amount of training samples.

6.1. Effect of sample size

The current deep learning models require a huge amount of training examples to achieve state-of-the-art performance. However, many application domains lack such a massive volume of training examples such as in certain medical imaging and behavioral analysis studies. Moreover, prospective acquisition of data may also be expensive in terms of both human and computing resources. The superior performance of deep models comes at the cost of network complexity, which is often hard to optimize and prone to overfitting without a large number of samples to train hundreds and thousands of parameters. Many research studies tend to present over-optimistic performance with deep models without proper validation or proof of generalization across datasets. Some of the solutions such as data augmentation [239,240], transfer learning [241], and introduction of Bayesian concepts [242,243] have laid the groundwork for using small data, which we expect to progress over time. The second potential future direction in deep learning research may involve improving the architectures to efficiently handle high dimensional imaging data. In medical imaging, cardiovascular imaging involves time-sampled 3D images of the heart as 4D data. The analysis of videos of 3D models and 3D point clouds is computationally intensive. Since the current deep CNN models are primarily designed to handle 2D images, the models are often extended to handle 3D volumes. This is accomplished by either converting the information to 2D sequences or utilizing dimensionality reduction techniques in the preprocessing stage. However, important information in volume data may be lost due to this conversion. Therefore, a carefully designed deep learning architecture that is capable of efficiently handling raw 3D data similar to their 2D counterparts is highly desirable.

6.2. Computational burden on mobile platforms

The computational expense is one of the major obstacles to using deep model in personal devices and making the technology as ubiquitous as the internet of things. Current state-of-the-art deep learning models utilize an enormous amount of hardware resources, which prohibit deploying them in most practical environments. As discussed in Sections 4.1–4.3, we believe that improvements in efficiency and memory footprints may enable the seamless utilization of mobile and wearable devices. An emerging deep learning research area involves achieving real-time learning in memory-constrained applications. Such real-time operation will require careful selection of learning models, model parameterization, and sophisticated hardware-software co-design.

6.3. Interpretability of models

The complexity in network architecture has been a critical factor in providing useful interpretations of model outcomes. In most applications, deep models are used as ‘black-box’ and optimized using heuristic methods for different tasks. For example, dropout has been introduced to combat model overfitting [242,244] to optimize the network performance. Dropout essentially deactivates a number of neurons at random without learning which neurons and weights are truly important. More importantly, the importance of input features and the inner working principles are not well understood in deep models. There has been limited progress in understanding the theoretical underpinning of these networks and more work is needed in this area.

6.4. Pitfalls of over-optimism

In a few applications such as in the game of GO, deep models have outperformed humans [245] and that has led to the notion that intelligent systems may replace human experts in the future. However, the vision-based intelligent algorithms may not be solely relied on for critical decision-making such as in clinical diagnosis without the supervision of a radiologist. While deep neural networks can perform many routine, repetitive, and predictive tasks better than human senses (such as vision) can offer, intelligent machines are unable to master many real-life inherently human-level traits such as empathy. Therefore, neural networks are developing intelligent systems that may be better viewed as complementary tools to assist and optimize human performance and decision-making.

7. Summary of survey

This paper systematically reviews the most recent progress and innovations of deep learning algorithms in speech and vision applications and limitations in implementating these models on popular mobile and embedded devices. The rapid evolution and success of deep learning algorithms is pioneering many new applications and commercial initiatives pertaining to intelligent speech and vision systems, which in turn is improving our daily lives. Despite tremendous success and performance gains of deep learning algorithms, there remain substantial challenges in implementing standalone speech and vision applications on mobile and resource constrained devices. Future research efforts will reach out to billions of mobile phone users with the most sophisticated deep learning-based intelligent systems. From sentiment and emotion recognition to developing self-driving intelligent transportation systems, there is a long list of speech and vision applications that will gradually automate and assist human’s visual and auditory perception to a greater scale and precision. With an overview of emerging applications across many disciplines such as behavioral science, psychology, transportation, and medicine, this paper serves as an excellent foundation for researchers, practitioners, and application developers and users.

The key observations for this survey paper are summarized below. First, we provide an overview of different state-of-the-art DNN algorithms and architectures in speech and vision applications. Several variants of CNN models [33,92–98] are proposed to address critical challenges related to vision-related recognition. Currently, CNN is one of the successful and dynamic areas of research and is dominating state-of-the-art vision systems both in the industry and academia. In addition, we briefly survey several other pioneering DNN architectures, such as DBNs, DBMs, GANs, VAEs, and SAEs in speech and vision recognition applications. RNN models are leading the current speech recognition systems, especially in the emerging applications of NLP. Several revolutionary variants of RNN such as the non-linear structure of LSTM [130,246] and the hybrid CNN-LSTM architecture [247] have made substantial improvements in the field of intelligent speech recognition and automatic image captioning.

Second, we address several challenges for state-of-the-art neural networks in adapting to compact and mobile platforms. Despite tremendous success in performance, the state-of-the-art intelligent algorithms entail heavy computation, memory usage, and power consumption. Studies on embedded intelligent systems, such as speech recognition and keyword spotting, are focused on adapting the most robust deep language models to resource restricted hardware available in mobile devices. Several studies [167–170,173] have customized DNN, CNN, and recurrent LSTM architectures with compression and quantization schemes to achieve considerable reductions in memory and computational requirements. Similarly, recent studies on embedded computer vision models suggest lightweight, efficient deep architectures [175,183,185] that are capable of real-time performance on existing mobile CPU and GPU hardware. We further identify several studies on developing computational algorithms and software systems [181,189,248] that augment the efficiency of contemporary deep models regardless of the recognition task. In addition, we identify the need for further research in developing effective learning algorithms for deep models using a minimal amount of training samples. Also, more computationally efficient architectures are expected to emerge to fully incorporate complex 3D/4D imaging data in learning. Moreover, fundamental research in hardware-software co-design is needed to address real-time learning operation for today’s memory-constrained cyber and physical systems.

Third, we identify three areas that are undergoing a paradigm shift largely driven by speech and vision-based intelligent systems. The vision or speech-based recognition of human emotion and behavior is revolutionizing a range of disciplines from behavioral science and psychology to consumer research and human–computer interactions. Intelligent applications for driver’s assistant and self-driving cars can greatly benefit from vision-based computational systems for future traffic management and driverless autonomous services. Deep neural networks in vision-based intelligent systems are rapidly transforming clinical research with the promise of futuristic precision diagnostic tools. Finally, we highlight three limitations of deep models: pitfalls of using small datasets, hardware constraints in mobile devices, and the danger of over-optimism to replace human experts by intelligent systems.

We hope this comprehensive survey in deep neural networks for speech and vision processing will serve as a key technical resource for future innovations and evolutions in autonomous systems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge partial funding of this work by the National Science Foundation (NSF), USA through a grant (Award# ECCS 1310353) and the National Institute of Health (NIH), USA through a grant (NIBIB/NIH grant# R01 EB020683). Note the views and findings reported in this work completely belong to the authors and not the NSF or NIH.

References

[1] Y. Dong, Z. Hu, K. Uchimura, N. Murayama, Driver inattention monitoring system for intelligent vehicles: a review, IEEE Trans. Intell. Transport. Syst. 12 (2) (2011) 596–614, https://doi.org/10.1109/TITS.2010.2092770.
[2] J.C. McCall, M.M. Trivedi, Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation, IEEE Trans. Intell. Transport. Syst. 7 (1) (2006) 20–37, https://doi.org/10.1109/TITS.2006.869595.
[3] N. Buch, S.A. Velastin, J. Orwell, A review of computer vision techniques for the analysis of urban traffic, IEEE Trans. Intell. Transp. Syst. 12 (3) (2011) 920–939, https://doi.org/10.1109/TITS.2011.2119372.
[4] E. Ohn-Bar, M.M. Trivedi, Looking at humans in the age of self-driving and highly automated vehicles, IEEE Trans. Intell. Veh. 1 (1) (2016) 90–104, https://doi.org/10.1109/TIV.2016.2571067.
[5] M. Bojarski et al., "End to End Learning for Self-Driving Cars," arXiv:1604, pp. 1-9, 2016. [Online]. Available: http://arxiv.org/abs/1604.07316.
[6] H. Woo et al., Lane-change detection based on vehicle-trajectory prediction, IEEE Rob. Autom. Lett. 2 (2) (2017) 1109–1116, https://doi.org/10.1109/LRA.2017.2660543.
[7] W. Ouyang, X. Zeng, X. Wang, Single-pedestrian detection aided by two-pedestrian detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1875–1889, https://doi.org/10.1109/TPAMI.2014.2377734.
[8] W. Huang, G. Song, H. Hong, K. Xie, Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning, IEEE Trans. Intell. Transp. Syst. 15 (5) (2014) 2191–2201, https://doi.org/10.1109/TITS.2014.2311123.
[9] X. Wang, R. Jiang, L. Li, Y. Lin, X. Zheng, F.-Y. Wang, Capturing Car-Following Behaviors by Deep Learning, IEEE Trans. Intell. Transp. Syst. (2017) 1–11, https://doi.org/10.1109/TITS.2017.2706963.
[10] A. Ferdowsi, U. Challita, and W. Saad, "Deep Learning for Reliable Mobile Edge Analytics in Intelligent Transportation Systems: An Overview," ieee vehicular technology magazine, vol. 14, no. 1, pp. 62-70, 2019.
[11] M. Havaei et al., Brain tumor segmentation with Deep Neural Networks, Med. Image Anal. 35 (2017) 18–31, https://doi.org/10.1016/j.media.2016.05.004.
[12] S. Liu et al., Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer's Disease, IEEE Trans. Biomed. Eng. 62 (4) (2015) 1132–1140, https://doi.org/10.1109/TBME.2014.2372011.
[13] E. Putin et al., Deep biomarkers of human aging: Application of deep neural networks to biomarker development, Aging 8 (5) (2016) 1021–1033, https://doi.org/10.18632/aging.100968.
[14] R. C. Deo et al., "An end-to-end computer vision pipeline for automated cardiac function assessment by echocardiography," CoRR, 2017.
[15] M.R. Alam, M.B.I. Reaz, M.A.M. Ali, "A review of smart homes—Past, present, and future," *IEEE Transactions on Systems, Man, and Cybernetics*, Part C (Applications and Reviews) 42 (6) (2012) 1190–1203.
[16] R. S. Cooper, J. F. McElroy, W. Rolandi, D. Sanders, R. M. Ulmer, and E. Peebles, "Personal virtual assistant," ed: Google Patents, 2011.
[17] E.W. Ngai, L. Xiu, D.C. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Syst. Appl. 36 (2) (2009) 2592–2602.
[18] S. Goswami, S. Chakraborty, S. Ghosh, A. Chakrabarti, B. Chakraborty, A review on application of data mining techniques to combat natural disasters, Ain Shams Eng. J. (2016) 1–14.
[19] S.S. Rautaray, A. Agrawal, Vision based hand gesture recognition for human computer interaction: a survey, Artif. Intell. Rev. 43 (1) (2015) 1–54.
[20] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.
[21] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in Advances in neural information processing systems, 2014, pp. 1799-1807.
[22] S. Srivastava, A. Bisht, and N. Narayan, "Safety and security in smart cities using artificial intelligence—A review," in Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on, 2017: IEEE, pp. 130-133.
[23] Y. Yang, C. Fairbairn, J.F. Cohn, Detecting depression severity from vocal prosody, IEEE Trans. Affective Comput. 4 (2) (2013) 142–150, https://doi.org/10.1109/T-AFFC.2012.38.
[24] L.D. Shriberg, R. Paul, J.L. McSweeny, A. Klin, D.J. Cohen, F.R. Volkmar, Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome, J. Speech, Language, Hearing Res. 44 (5) (2001) 1097–1115, https://doi.org/10.1044/1092-4388(2001/087).
[25] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recogn. 44 (3) (2011) 572–587, https://doi.org/10.1016/j.patcog.2010.09.020.
[26] H.M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for Speech Emotion Recognition, Neural Networks 92 (2017) 60–68, https://doi.org/10.1016/j.neunet.2017.02.013.
[27] Y. Kim, H. Lee, E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," 2013, pp. 3687-3691, doi: 10.1109/ICASSP.2013.6638346. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6638346.
[28] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[29] G.E. Hinton, Learning multiple layers of representation, Trends Cognitive Sci. 11 (10) (2007) 428–434.
[30] R.M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence, Sci. Rep. 6 (2016) 1–13, Art no. 27755.
[31] N. Kruger et al., Deep hierarchies in the primate visual cortex: What can we learn for computer vision?, IEEE Trans Pattern Anal. Mach. Intell. 35 (8) (2013) 1847–1871.
[32] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Networks 61 (2015) 85–117.
[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097-1105.
[34] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (Dec) (2010) 3371–3408.
[35] I. Goodfellow et al., Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672-2680.
[36] D.P. Kingma, M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.
[37] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," arXiv preprint arXiv:1605.08803, 2016.
[38] Z. C. Lipton, J. Berkowitz, C. Elkan, "A critical review of recurrent neural networks for sequence learning," arXiv preprint arXiv:1506.00019, pp. 1-38, 2015.
[39] A. Vaswani et al., Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998-6008.
[40] M. Alam, L. Vidyaratne, K.M. Iftekharuddin, Novel hierarchical Cellular Simultaneous Recurrent neural Network for object detection, in: Neural Networks (IJCNN), 2015 International Joint Conference on, 12-17 July 2015 2015, pp. 1-7, doi: 10.1109/IJCNN.2015.7280480.
[41] R. Salakhutdinov, A. Mnih, G. Hinton, Restricted Boltzmann machines for collaborative filtering, in: Proceedings of the 24th international conference on Machine learning, 2007: ACM, pp. 791-798.
[42] R. Salakhutdinov, G. Hinton, Deep boltzmann machines, Artificial Intelligence Statistics (2009) 448–455.
[43] J. Gehring, Y. Miao, F. Metze, A. Waibel, Extracting deep bottleneck features using stacked auto-encoders, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013: IEEE, pp. 3377-3381.
[44] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th international conference on Machine learning, 2008: ACM, pp. 1096-1103.
[45] G. B. Huang, H. Lee, E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012: IEEE, pp. 2518-2525.
[46] Z. You, X. Wang, and B. Xu, "Investigation of deep boltzmann machines for phone recognition," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: IEEE, pp. 7600-7603.
[47] G. Wen, H. Li, J. Huang, D. Li, E. Xun, Random deep belief networks for recognizing emotions from speech signals, Comput. Intelligence Neuroscience 2017 (2017) 1–9.
[48] C. Huang, W. Gong, W. Fu, D. Feng, A research of speech emotion recognition based on deep belief network and SVM, Math. Problems Eng. 2014 (2014) 1–7.

[49] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in: Proceedings of the 26th annual international conference on machine learning, 2009: ACM, pp. 609-616.

[50] X. Yan, J. Yang, K. Sohn, H. Lee, Attribute2image: Conditional image generation from visual attributes, in: European Conference on Computer Vision, Springer, 2016, pp. 776–791.

[51] J. Walker, C. Doersch, A. Gupta, M. Hebert, An uncertain future: Forecasting from static images using variational autoencoders, in: European Conference on Computer Vision, Springer, 2016, pp. 835–851.

[52] S. Semeniuta, A. Severyn, E. Barth, "A hybrid convolutional variational autoencoder for text generation," arXiv preprint arXiv:1702.02390, 2017.

[53] K. Akuzawa, Y. Iwasawa, Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," arXiv preprint arXiv:1804.02135, 2018.

[54] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, "Generative adversarial text to image synthesis," arXiv preprint arXiv:1605.05396, 2016.

[55] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," arXiv preprint, 2017

[56] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.

[57] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.

[58] J. Donahue, P. Krähenbühl, T. Darrell, "Adversarial feature learning," arXiv preprint arXiv:1605.09782, 2016.

[59] J. Donahue, K. Simonyan, "Large scale adversarial representation learning," in: Advances in Neural Information Processing Systems, 2019, pp. 10541-10551

[60] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," arXiv preprint arXiv:1701.04862, 2017.

[61] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," arXiv preprint arXiv:1701.00160, 2016.

[62] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," arXiv preprint arXiv:1802.05957, 2018.

[63] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International conference on machine learning, 2017, pp. 214–223.

[64] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, "Improved training of wasserstein gans," in: Advances in neural information processing systems, 2017, pp. 5767-5777.

[65] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.

[66] A. Razavi, A. v. d. Oord, O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," arXiv preprint arXiv:1906.00446, 2019.

[67] L. Dinh, D. Krueger, Y. Bengio, "Nice: Non-linear independent components estimation," arXiv preprint arXiv:1410.8516, 2014.

[68] D.P. Kingma, P. Dhariwal, Glow generative flow with invertible 1x1 convolutions, Adv. Neural Inform. Process. Syst. (2018) 10215–10224.

[69] A. v. d. Oord et al., "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[70] A. v. d. Oord, N. Kalchbrenner, K. Kavukcuoglu, "Pixel recurrent neural networks," arXiv preprint arXiv:1601.06759, 2016.

[71] R. Prenger, R. Valle, B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: IEEE, pp. 3617-3621.

[72] S. Pascual, A. Bonafonte, J. Serra, "SEGAN: Speech enhancement generative adversarial network," arXiv preprint arXiv:1703.09452, 2017

[73] N. Adiga, Y. Pantazis, V. Tsiaras, Y. Stylianou, Speech Enhancement for Noise-Robust Speech Synthesis Using Wasserstein GAN, Proc. Interspeech 2019 (2019) 1821–1825.

[74] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," arXiv preprint arXiv:1603.01354, 2016.

[75] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259, 2014

[76] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, IEEE Trans. Neural Networks Learn. Syst. 28 (10) (2016) 2222–2232.

[77] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in Advances in neural information processing systems, 2014, pp. 2204-2212.

[78] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in Advances in neural information processing systems, 2010, pp. 1243-1251.

[79] M. A. Ranzato, "On learning where to look," arXiv preprint arXiv:1405.5488, 2014.

[80] M. Denil, L. Bazzani, H. Larochelle, N. de Freitas, Learning where to attend with deep architectures for image tracking, Neural Comput. 24 (8) (2012) 2151–2184.

[81] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," arXiv preprint arXiv:1502.04623, 2015

[82] K. Fu, J. Jin, R. Cui, F. Sha, C. Zhang, Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2321–2334, https://doi.org/10.1109/TPAMI.2016.2642953.

[83] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," arXiv preprint arXiv:1511.02793, 2015.

[84] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," arXiv preprint arXiv:1410.5401, pp. 1-26, 2014.

[85] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, pp. 1-11, 2015

[86] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, 2016: IEEE, pp. 4960-4964.

[87] J. Liu, A. Shahroudy, D. Xu, A.C. Kot, G. Wang, Skeleton-based action recognition using spatio-temporal LSTM network with trust gates, IEEE Trans. Pattern Anal. Mach. Intell. 40 (12) (2018) 3007–3021.

[88] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," arXiv preprint arXiv:1805.08318, 2018.

[89] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," arXiv preprint arXiv:1611.01578, 2016.

[90] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," arXiv preprint arXiv:1806.09055, 2018.

[91] C. Liu et al., Progressive neural architecture search, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–34.

[92] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[93] C. Szegedy et al., Going deeper with convolutions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[94] O. Russakovsky et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, pp. 1-14, 2014

[96] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[97] Z. Wu, C. Shen, and A. v. d. Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," arXiv preprint arXiv:1611.10080, pp. 1-19, 2016

[98] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[99] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[100] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[101] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946, 2019.

[102] H. Wang et al., Cosface: Large margin cosine loss for deep face recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[103] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[104] P. Wang, Y. Cao, C. Shen, L. Liu, H.T. Shen, Temporal pyramid pooling based convolutional neural networks for action recognition, IEEE Trans. Circuits Syst. Video Technol. 27 (12) (2017) 2613–2622.

[105] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1080–1088.

[106] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.

[107] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, S.A. Velastin, Exploiting deep residual networks for human action recognition from skeletal data, Comput. Vis. Image Underst. 170 (2018) 51–66.

[108] Y. Tang, Y. Tian, J. Lu, P. Li, J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5323–5332.

[109] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, D.D. Feng, Deep convolutional neural networks for human action recognition using depth maps and postures, IEEE Trans. Syst. Man Cybern, Syst. (2018).

[110] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vision 61 (1) (2005) 55–79.

[111] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, X. Wang, 3d human pose estimation in the wild by adversarial learning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5255–5264.

[112] L. Ge, H. Liang, J. Yuan, D. Thalmann, Real-time 3D hand pose estimation with 3D convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2019) 956–970.

[113] R. Alp Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[114] T. Wang et al., Detect globally, refine locally: A novel approach to saliency detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3127–3135.

[115] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 714–722.
[116] Z. Wang, J. Ren, D. Zhang, M. Sun, J. Jiang, A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos, Neurocomputing 287 (2018) 68–83.
[117] H. Song, W. Wang, S. Zhao, J. Shen, K.-M. Lam, Pyramid dilated deeper convlstm for video salient object detection, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 715–731.
[118] L. Leal-Taixé, C. Canton-Ferrer, K. Schindler, Learning by tracking: Siamese CNN for robust target association, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 33–40.
[119] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P.H. Torr, Fast online object tracking and segmentation: a unifying approach, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.
[120] T. Dai et al., "Deep Reinforcement Learning for Subpixel Neural Tracking," in International Conference on Medical Imaging with Deep Learning, 2019, pp. 130-150.
[121] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
[122] R.A. Yeh, C. Chen, T. Yian Lim, A.G. Schwing, M. Hasegawa-Johnson, M.N. Do, Semantic image inpainting with deep generative models, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.
[123] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.
[124] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.
[125] M. Kumar et al., "VideoFlow: A flow-based generative model for video," arXiv preprint arXiv:1903.01434, 2019.
[126] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černocký, "Strategies for training large scale neural network language models," in Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, 2011: IEEE, pp. 196-201.
[127] G. Hinton et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Process Mag. 29 (6) (2012) 82–97.
[128] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013: IEEE, pp. 8614-8618.
[129] H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in: *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 338–342.
[130] J.-T. Chien and A. Misbullah, "Deep long short-term memory networks for speech recognition," in: Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on, 2016: IEEE, pp. 1-5.
[131] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018: IEEE, pp. 5934-5938.
[132] C.-C. Chiu et al., "State-of-the-art speech recognition with sequence-to-sequence models," in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: IEEE, pp. 4774-4778.
[133] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," arXiv preprint arXiv:1805.03294, 2018.
[134] J. Weston, S. Chopra, and A. Bordes, "Memory networks," arXiv preprint arXiv:1410.3916, pp. 1-15, 2014
[135] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," arXiv preprint arXiv:1503.00075, pp. 1-11, 2015.
[136] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," arXiv preprint arXiv:1609.08144, pp. 1-23, 2016
[137] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
[138] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: IEEE, pp. 2227-2231.
[139] M. Chen, X. He, J. Yang, H. Zhang, 3-D convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process Lett. 25 (10) (2018) 1440–1444.
[140] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," arXiv preprint arXiv:1806.02146, 2018.
[141] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2018).
[142] T. Stafylakis, G. Tzimiropoulos, Zero-shot keyword spotting for visual speech recognition in-the-wild, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 513–529.
[143] A. Krizhevsky, V. Nair, and G. Hinton, "The CIFAR-10 dataset," online: http:// www. cs. toronto. edu/kriz/cifar. html, vol. 55, 2014.
[144] T.-Y. Lin et al., Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[145] D. Du et al., The unmanned aerial vehicle benchmark: Object detection and tracking, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.
[146] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," arXiv preprint arXiv:1804.07437, 2018.
[147] C. Lopes, F. Perdigao, Phone recognition on the TIMIT database, Speech Technologies/Book 1 (2011) 285–302.
[148] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.
[149] Stanford. "Neural Machine Translation." https://nlp.stanford.edu/projects/nmt/ (accessed.
[150] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth'CHiME'Speech Separation and Recognition Challenge: Dataset, task and baselines," arXiv preprint arXiv:1803.10609, 2018.
[151] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," arXiv preprint arXiv:1809.00496, 2018.
[152] Google. "Google Brain Team's Mission." https://ai.google/research/teams/brain/ (accessed.
[153] Facebook. "Facebook AI Research (FAIR)." https://research.fb.com/category/facebook-ai-research-fair/ (accessed.
[154] T. Simonite, "Facebook's Perfect, Impossible Chatbot," MIT Technology Review. [Online]. Available: https://www.technologyreview.com/s/604117/facebooks-perfect-impossible-chatbot/
[155] Microsoft. "Cognitive Toolkit." https://docs.microsoft.com/en-us/cognitive-toolkit/ (accessed.
[156] W. Xiong et al., "Achieving human parity in conversational speech recognition," arXiv preprint arXiv:1610.05256, pp. 1-13, 2016.
[157] Microsoft. "Cortana." https://www.microsoft.com/en-us/cortana (accessed.
[158] I. T. Association, "Specification FAQ." [Online]. Available: http://www.infinibandta.org/content/pages.php?pg=technology_faq
[159] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in International Conference on Machine Learning, 2016, pp. 173-182.
[160] NVIDIA. "Deep Learning AI." https://www.nvidia.com/en-us/deep-learning-ai/ (accessed.
[161] IBM. "Watson." https://www.ibm.com/watson/ (accessed.
[162] A. Inc. "Apple Machine Learning Journal." https://machinelearning.apple.com/ (accessed.
[163] A. W. Services. "Amazon Machine Learning." https://aws.amazon.com/sagemaker (accessed.
[164] U. Engineering, "Engineering More Reliable Transportation with Machine Learning and AI at Uber." [Online]. Available: https://eng.uber.com/machine-learning/
[165] Intel, "Machine Learning Offers a Path to Deeper Insight." [Online]. Available: https://www.intel.com/content/www/us/en/analytics/machine-learning/overview.html
[166] J. Schalkwyk et al., "Your Word is my Command": Google Search by Voice: A Case Study, in: Advances in Speech Recognition, Springer, 2010, pp. 61–90.
[167] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014: IEEE, pp. 4087-4091.
[168] T.N. Sainath, C. Parada, Convolutional neural networks for small-footprint keyword spotting, in: *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 1478–1482.
[169] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015: IEEE, pp. 5236-5240.
[170] X. Lei, A.W. Senior, A. Gruenstein, J. Sorensen, Accurate and compact large vocabulary speech recognition on mobile devices, Interspeech 1 (2013) 662–665.
[171] B. Ballinger, C. Allauzen, A. Gruenstein, J. Schalkwyk, On-demand language model interpolation for mobile speech input, Interspeech (2010) 1812–1815.
[172] J. Sorensen, C. Allauzen, Unary data structures for language models, in: *Twelfth Annual Conference of the International Speech Communication Association*, 2011, pp. 1425–1428.
[173] Y. Wang, J. Li, and Y. Gong, "Small-footprint high-performance deep neural network-based speech recognition using split-VQ," in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015: IEEE, pp. 4984-4988.
[174] G. Tucker, M. Wu, M. Sun, S. Panchapagesan, G. Fu, S. Vitaladevuni, Model compression applied to small-footprint keyword spotting, INTERSPEECH (2016) 1878–1882.
[175] S. Sarkar, V. M. Patel, and R. Chellappa, "Deep feature-based face detection on mobile devices," in: Identity, Security and Behavior Analysis (ISBA), 2016 IEEE International Conference on, 2016: IEEE, pp. 1-8.
[176] Y. Bengio et al., Deep learners benefit more from out-of-distribution examples, in: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 164–172.

[177] M. E. Fathy, V. M. Patel, and R. Chellappa, "Face-based active authentication on mobile devices," in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015: IEEE, pp. 1687-1691.
[178] C. McCool and S. Marcel, "Mobio database for the ICPR 2010 face and speech competition," Idiap, 2009.
[179] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[180] J. Su et al., "Redundancy-Reduced MobileNet Acceleration on Reconfigurable Logic for ImageNet Classification," Cham, 2018: Springer International Publishing, in Applied Reconfigurable Computing. Architectures, Tools, and Applications, pp. 16-28.
[181] S. Han, H. Mao, W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, pp. 1-14, 2015.
[182] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," arXiv preprint arXiv:1606.06160, 2016.
[183] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, and F. Kawsar, "An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices," in: Proceedings of the 2015 International Workshop on Internet of Things towards Applications, 2015: ACM, pp. 7-12.
[184] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," arXiv preprint arXiv:1312.6082, pp. 1-13, 2013.
[185] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?," in: Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications, 2015: ACM, pp. 117-122.
[186] N. D. Lane et al., "Deepx: A software accelerator for low-power deep learning inference on mobile devices," in: Information Processing in Sensor Networks (IPSN), 2016 15th ACM/IEEE International Conference on, 2016: IEEE, pp. 1-12.
[187] N. Evans, Z. Wu, J. Yamagishi, T. Kinnunen, "Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database," 2015.
[188] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, NIPS workshop on deep learning and unsupervised feature learning 2011 (2) (2011) 5.
[189] V. Sindhwani, T. Sainath, S. Kumar, Structured transforms for small-footprint deep learning, Adv. Neural Inform. Process. Syst. (2015) 3088–3096.
[190] V. Pan, Structured matrices and polynomials: unified superfast algorithms. Springer Science & Business Media, 2012
[191] S. Wang and J. Jiang, "Learning natural language inference with LSTM," arXiv preprint arXiv:1512.08849, pp. 1-10, 2015
[192] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.
[193] N. Check. "Qualcomm Snapdragon 865." https://www.notebookcheck.net/Qualcomm-Snapdragon-865-SoC-Benchmarks-and-Specs.448194.0.html (accessed.
[194] Qualcomm. "Mobile Artificial Intelligence." https://www.qualcomm.com/products/smartphones/mobile-ai (accessed.
[195] N. Check. "Apple A13 Bionic vs Qualcomm Snapdragon 855+ / 855 Plus vs Apple A12 Bionic." (accessed.
[196] T. Centurion. "Best Mobile Processor Ranking List 2020." https://www.techcenturion.com/smartphone-processors-ranking (accessed.
[197] Apple. "Core ML Framework." https://developer.apple.com/documentation/coreml (accessed.
[198] A. Insider. "Why the Apple A13 Bionic blows past Qualcomm Snapdragon 855 Plus." https://appleinsider.com/articles/19/10/22/editorial-why-the-apple-a13-bionic-blows-past-qualcomm-snapdragon-855-plus (accessed.
[199] W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge Computing: Vision and Challenges, IEEE Internet Things J. 3 (5) (2016) 637–646, https://doi.org/10.1109/JIOT.2016.2579198.
[200] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, D. Sabella, On multi-access edge computing: a survey of the emerging 5g network edge cloud architecture and orchestration, IEEE Commun. Surv. Tutorials 19 (3) (2017) 1657–1681, https://doi.org/10.1109/COMST.2017.2705720.
[201] S. Yi, Z. Hao, Z. Qin, Q. Li, "Fog computing: Platform and applications," in: 2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb), 2015: IEEE, pp. 73-78.
[202] K. Ha, Z. Chen, W. Hu, W. Richter, P. Pillai, M. Satyanarayanan, in: Towards wearable cognitive assistance, and services, 2014, pp. 68–81.
[203] K. Kumar, Y. Lu, Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?, Computer 43 (4) (2010) 51–56, https://doiorg/10.1109/MC.2010.98.
[204] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, A. Patti, Clonecloud: elastic execution between mobile device and cloud, in: *Proceedings of the sixth conference on Computer systems*, 2011, pp. 301–314.
[205] S. Wu, C. Mei, H. Jin, D. Wang, Android Unikernel: Gearing mobile code offloading towards edge computing, Future Generation Computer Systems 86 (2018,) 694–703, https://doi.org/10.1016/j.future.2018.04.069.
[206] Intel. "Intel® Neural Compute Stick 2." https://software.intel.com/en-us/neural-compute-stick (accessed April 2, 2020).
[207] NVIDIA. "NIVIDIA Jetson." https://developer.nvidia.com/buy-jetson (accessed April 2, 2020).
[208] R.A. Calvo, S. D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, IEEE Trans. Affective Comput. 1 (1) (2010) 18–37, https://doi.org/10.1109/T-AFFC.2010.1.
[209] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, vol. 2: IEEE, pp. 568-573.
[210] E. M. Albornoz, M. Sánchez-Gutiérrez, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Spoken Emotion Recognition Using Deep Learning," Springer, Cham, 2014, pp. 104-111
[211] S. Wang and Q. Ji, "Video affective content analysis: a survey of state of the art methods," IEEE Transactions on Affective Computing, vol. 6, no. 4, pp. 1-1, 2015, doi: 10.1109/TAFFC.2015.2432791.
[212] M. G. Ball, B. Qela, and S. Wesolkowski, "A review of the use of computational intelligence in the design of military surveillance networks," in: Recent Advances in Computational Intelligence in Defense and Security: Springer, 2016, pp. 663-693.
[213] R. Olmos, S. Tabik, F. Herrera, Automatic handgun detection alarm in videos using deep learning, Neurocomputing 275 (2018) 66–72.
[214] X. Li et al., Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods, IEEE Trans. Affective Comput. (2017).
[215] P. Ekman, Friesen, W. V., Hager, J. C, "Facial Action Coding System - Manual and Investigator's Guide. FACS," Research Nexus, 2002. [Online]. Available: https://doi.org/10.1016/j.msea.2004.04.064.
[216] J. Whitehill, Z. Serpell, Y.C. Lin, A. Foster, J.R. Movellan, The faces of engagement: Automatic recognition of student engagement from facial expressions, IEEE Trans. Affective Comput. 5 (1) (2014) 86–98, https://doi.org/10.1109/TAFFC.2014.2316163.
[217] K.A. Leitch, S.E. Duncan, S. O'Keefe, R. Rudd, D.L. Gallagher, Characterizing consumer emotional response to sweeteners using an emotion terminology questionnaire and facial expression analysis, Food Res. Int. 76 (2015) 283–292, https://doi.org/10.1016/j.foodres.2015.04.039.
[218] C.F. Camerer, Artificial intelligence and behavioral economics, Economics of Artificial Intelligence, University of Chicago Press, 2017.
[219] M.D. Samad, N. Diawara, J.L. Bobzien, J.W. Harrington, M.A. Witherow, K.M. Iftekharuddin, A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (2) (2018) 353–361.
[220] M. Leo et al., Computational analysis of deep visual data for quantifying facial expression production, Appl. Sci. 9 (21) (2019) 4542.
[221] M.D. Samad, N. Diawara, J.L. Bobzien, C.M. Taylor, J.W. Harrington, K.M. Iftekharuddin, A pilot study to identify autism related traits in spontaneous facial actions using computer vision, Res. Autism Spectrum Disorders 65 (2019) 14–24.
[222] Audi. "Autonomous Driving." https://www.audi.com/en/experience-audi/mobility-and-trends/autonomous-driving.html (accessed.
[223] Tesla. "All Tesla Cars Being Produced Now Have Full Self-Driving Hardware." https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware (accessed.
[224] C. Wang, Y. Hou, M. Barth, Data-Driven Multi-step Demand Prediction for Ride-Hailing Services Using Convolutional Neural Network, in: Science and Information Conference, Springer, 2019, pp. 11–22.
[225] S. Das et al., "Map Enhanced Route Travel Time Prediction using Deep Neural Networks," arXiv preprint arXiv:1911.02623, 2019.
[226] A. Alabbasi, A. Ghosh, and V. Aggarwal, "Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning," arXiv preprint arXiv:1903.03882, 2019.
[227] M. Daniel, M.A. Makary, Medical error—the third leading cause of death in the US, BMJ 353 (i2139) (2016) 476636183.
[228] A. Ulloa et al., "A deep neural network predicts survival after heart imaging better than cardiologists," arXiv preprint arXiv:1811.10553, 2018.
[229] A. Ulloa et al., "A deep neural network to enhance prediction of 1-year mortality using echocardiographic videos of the heart," arXiv preprint arXiv:1811.10553, 2018.
[230] E. Rahimy, Deep learning applications in ophthalmology, Curr. Opin. Ophthalmol. 29 (3) (2018) 254–260.
[231] J.-H. Lee, D.-H. Kim, S.-N. Jeong, S.-H. Choi, Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm, J. Dent. 77 (2018) 106–111.
[232] A. Esteva et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115.
[233] E. Gong, J.M. Pauly, M. Wintermark, G. Zaharchuk, Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI, J. Magn. Reson. Imaging 48 (2) (2018) 330–340.
[234] G.A. Bello et al., Deep-learning cardiac motion analysis for human survival prediction, Nature Mach. Intelligence 1 (2) (2019) 95.
[235] O. Bernard et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?, IEEE Trans Med. Imaging 37 (11) (2018) 2514–2525.
[236] E. Arvaniti et al., "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," Scientific reports, vol. 8, 2018.
[237] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, S. Mougiakakou, Lung pattern classification for interstitial lung diseases using a deep

convolutional neural network, IEEE Trans. Med. Imaging 35 (5) (2016) 1207–1216.

[238] M.R. Arbabshirani et al., Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration, npj Digital Med. 1 (1) (2018) 9.

[239] C. C. Charalambous and A. A. Bharath, "A data augmentation methodology for training machine/deep learning gait recognition algorithms," arXiv preprint arXiv:1610.07570, pp. 1-12, 2016

[240] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," arXiv preprint arXiv:1609.08764, pp. 1-6, 2016

[241] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: a survey, Knowl.-Based Syst. 80 (2015) 14–23.

[242] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: *international conference on machine learning*, 2016, pp. 1050–1059.

[243] H. Wang and D.-Y. Yeung, "Towards bayesian deep learning: A survey," arXiv preprint arXiv:1604.01662, pp. 1-17, 2016

[244] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.

[245] D. Silver et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484.

[246] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

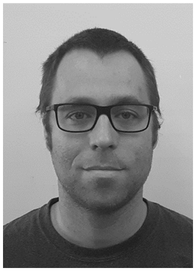
[247] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[248] A. Rakotomamonjy, G. Gasso, Histogram of gradients of time–frequency representations for audio scene classification, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1) (2014) 142–153.

Dr. Mahbubul Alam completed his Ph.D. from the Vision Lab of the Department of Electrical and Computer Engineering at the Old Dominion University (ODU), Norfolk, Virginia, USA in December 2018. His doctoral dissertation contributed to developing novel machine learning, more specifically deep learning algorithms for solving complex computer vision problems. After finishing his Ph.D., he started working as a Research Scientist at the Industrial AI Lab, Research and Development Division at Hitachi America, Ltd., Santa Clara, California. Dr. Alam received his BS and MS degree in Computer Science and Engineering from Jahangirnagar University, Bangladesh in 2008 and 2011, respectively. He has secured highest grade in both BS and MS final examination. He was awarded "Gold Medal" for obtaining the highest Cumulative GPA in the whole University during his undergrad study. His research interests are machine learning, deep learning, computer vision, predictive maintenance, visual inspection and other interesting and challenging problems in the industrial domain. He has served as the reviewer for the following IEEE journals: IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Emerging Topics in Computational Intelligence, and IEEE Journal of Biomedical and Health Informatics.

Dr.Manar D. Samad received the B.S. degree in Electrical and Electronic Engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2007, the M.S. degree in Computer Engineering from the University of Calgary, Calgary, AB, Canada in 2011, and the Ph.D. degree from Old Dominion University, Norfolk, VA, USA in 2016. He worked as a post-doctoral research fellow at Geisinger, Danville, PA, USA with the Department of Imaging Science and Innovation. He is currently an Assistant Professor in the Department of Computer Science at Tennessee State University, Nashville, TN, USA. His research interests include machine learning, health informatics, computer vision, and human computer interactions.

Dr.Lasitha S. Vidyaratne received the M.Eng. degree in Electronic and Communications Engineering from the University of Nottingham Malaysia Campus, Malaysia. He received the Ph.D. degree with the Vision Laboratory, Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, USA in 2019. His doctoral dissertation focuses on developing novel deep recurrent neural network architectures for complex computer vision and biomedical signal processing applications. His current research interests include deep learning, recurrent neural networks, unsupervised learning, reinforcement learning, brain-computer interfacing, and biomedical signal processing.



Alexander M. Glandon is currently a research assistant at Old Dominion University Vision Lab, Norfolk VA, USA. He is pursuing a PhD in Electrical and Computer Engineering at ODU, where he also received his M.S. in Electrical and Computer Engineering and B.S. in Computer Engineering. His current research analyzes 3D lidar depth maps of humans for identification and activity recognition using image processing and novel algorithms. He is also interested in machine learning and statistics.

Dr. Khan M. Iftekharuddin is a professor in the department of electrical and computer engineering and an Associate Dean for Research in the Batten College of Engineering at Old Dominion University (ODU). He is a Batten Endowed Chair in Machine Learning. He serves as the director of ODU Vision Lab. He is also a member of biomedical engineering program at ODU. Much of his research had focused on different aspects of computer vision, machine learning and signal/image processing problems. His research interests include machine learning and computational modeling, stochastic medical image analysis for tumor phenotype, intersection of bioinformatics and medical image analysis, automatic target recognition, biologically inspired human and machine centric learning, deep recurrent learning and networks, sensor signal acquisition and modeling, and optical computing and interconnection. He is the principal author of a book, several book chapters, and about 200 refereed journal and conference papers. His research has been funded by different agencies such as NSF, NIH, AFOSR, ARO, AFRL, ONR, DOT, DOE, Whitaker Foundation, Assisi Foundation, and Fed-EX-FIT. He obtained his B.Sc. degree from Bangladesh Institute of Technology in 1989. He received an M.S. and a Ph.D. both in electrical engineering from the University of Dayton in 1991 and 1995 respectively. He serves as a Senior Associate Editor for *Optical Engineering* journal and as an Associate Editor for multiple journals including *IEEE Transaction in Image Processing, IEEE Transaction in Learning Systems, Cognitive Systems Research* and *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. He is a Fellow of SPIE, a Senior Member of IEEE, INNS and OSA.