Multimodal Intelligence: Representation Learning, Information Fusion, and Applications

Chao Zhang, Zichao Yang, Xiaodong He, *Fellow, IEEE*, and Li Deng, *Fellow, IEEE*

***Abstract*—Deep learning methods haverevolutionized speech recognition, image recognition, and natural language processing since 2010. Each of these tasks involves a single modality in their input signals. However, many applications in the artificial intelligence field involve multiple modalities. Therefore, it is of broad interest to study the more difficult and complex problem of modeling and learning across multiple modalities. In this paper, we provide a technical review of available models and learning methods for multimodal intelligence. The main focus of this review is the combination of vision and natural language modalities, which has become an important topic in both the computer vision and natural language processing research communities. This review provides a comprehensive analysis of recent works on multimodal deep learning from three perspectives: learning multimodal representations, fusing multimodal signals at various levels, and multimodal applications. Regarding multimodal representation learning, we review the key concepts of embedding, which unify multimodal signals into a single vector space and thereby enable cross-modality signal processing. We also review the properties of many types of embeddings that are constructed and learned for general downstream tasks. Regarding multimodal fusion, this review focuses on special architectures for the integration of representations of unimodal signals for a particular task. Regarding applications, selected areas of a broad interest in the current literature are covered, including image-to-text caption generation, text-to-image generation, and visual question answering. We believe that this review will facilitate future studies in the emerging field of multimodal intelligence for related communities.**

***Index Terms*—Multimodality, representation, multimodal fusion, deep learning, embedding, speech, vision, natural language, caption generation, text-to-image generation, visual question answering, visual reasoning.**

I. Introduction

SIGNIFICANT progress has been made in the field of machine learning in recent years based on the rapid development of deep learning algorithms [1]–[6]. The first major milestone was a significant increase in the accuracy of large-scale automatic speech recognition based. on the use of fully connected deep neural networks (DNNs) and deep autoencoders around 2010 [7]–[17]. Shortly thereafter, a series of breakthroughs was achieved in computer vision (CV) using deep convolutional neural network (CNN) models [18] for large-scale image classification around 2012 [19]–[22] and large-scale object detection around 2014 [23]–[25]. All of these milestones have been achieved for pattern recognition with a single input modality. In natural language processing (NLP), recurrent neural network (RNN) based semantic slot filling methods [26] have achieved state-of-the-art for spoken language understanding. RNN-encoder-decoder models with attention mechanisms [27], which are also referred to as sequence-to-sequence models [28], have achieved superior performance for machine translation in an end-to-end fashion [29], [30]. For additional NLP tasks with small amounts of training data, such as question answering (QA) and machine reading comprehension, generative pre-training has achieved state-of-the-art results [31]–[33]. This method transfers parameters from a language model (LM) pre-trained on a large out-of-domain dataset using unsupervised training or self-training, which is followed by fine-tuning on small in-domain datasets.

Although there have been significant advances in vision, speech, and language processing, many problems in the artificial intelligence field involve more than one input modality, such as intelligent personal assistant systems that must understand human communication based on spoken words, body language, and pictorial languages [34]. Therefore, it is of broad interest to study modeling and training approaches across multiple modalities [35]. Based on advances in image processing and language understanding [36], tasks combining images and text have attracted significant attention, including visual-based referred expression understanding and phrase localization [37]–[39], as well as image and video captioning [40]–[45], visual QA (VQA) [46]–[48], text-to-image generation [49]–[51], and visual-and-language navigation [52]. In these tasks, natural language plays a key role in helping machines in "understanding" the content of images, where "understanding" means capturing the underlying correlations between the semantics embedded in languages and the visual features obtained from images. In addition to text, vision can also be combined with speech to perform audio-visual speech recognition [53]–[55], speaker recognition [56]–[58], speaker diarization, [59], [60], as well as speech separation [61], [62] and enhancement [63].

This paper provides a technical review of the models and training methods used for multimodal intelligence. Our main focus

Manuscript received November 10, 2019; revised March 23, 2020 and April 3, 2020; accepted April 7, 2020. Date of publication April 15, 2020; date of current version June 24, 2020. This work was supported by the Beijing Academy of Artificial Intelligence. The guest editor coordinating the review of this paper and approving it for publication was Dr. Isabel Trancoso. *(Corresponding author: Xiaodong He.)*

Chao Zhang was with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K., and also with the JD AI Research, JD.com Inc, Beijing 100101, China (e-mail: cz277@cam.ac.uk).

Zichao Yang was with the Citadel LLC, Chicago, IL 60603 USA (e-mail: yangtze2301@gmail.com).

Li Deng was with the Citadel America, Seattle, WA, 98121 USA (e-mail: deng629@gmail.com).

Xiaodong He was with the JD AI Research, JD.com Inc, Beijing 100101, China (e-mail: xiaodong.he@jd.com).

Digital Object Identifier 10.1109/JSTSP.2020.2987728

is the combination of CV and NLP, which has become an important area for both of these research communities that covers many different tasks and technologies. To provide a structured perspective, we have organized this technical review according to three key topics: representations, fusion, and applications.

- Learning representations for input data is a core problem in deep learning. For multimodal tasks, collecting parallel data across different modalities can be a difficult task. Leveraging pre-trained representations with the desired properties, such as properties suitable for zero-shot or few-shot learning, is often an effective solution to this issue. Both supervised and unsupervised training-based multimodal representation learning methods are reviewed.
- Fusing the representations of different modalities is a core problem in any multimodal task. Unlike previous studies that have classified related work based on the stage in which fusion occurs within a procedure, we classify related work according to the actual operations used during fusion, such as attention mechanisms and bilinear pooling, because it is difficult to classify recent complex approaches based on stages.
- Three types of applications are reviewed: image captioning, text-to-image generation, and VQA. These applications provide examples of how representation learning and fusion can be applied to specific tasks, as well as a representation of the current development of multimodal applications, particularly those integrating vision with natural languages. Visual reasoning methods are also discussed.

The remainder of this paper is organized as follows. Section II reviews recent progress in terms of developing representations for single or multiple modalities. Section III introduces commonly used fusion methods with a focus on attention mechanisms and bilinear pooling. Section IV introduces applications, including caption generation, text-to-image generation, VQA, and visual reasoning, followed by a summary and our outlook regarding potential future research directions.

II. Representations

Deep learning, as a special area within representational learning, focuses on the use of artificial neural networks (ANN) with many hidden layers to discover suitable representations or features from raw data automatically for specific tasks [64]. Representation learning has great value in practice since better representations can often simplify subsequent learning tasks. Over the past decade, it has become feasible to learn effective and robust representations for single modalities, such as text [31]–[33], [65]–[72] or images [19]–[25], based on the availability of large amounts of data and the development of deep learning. Although multimodal representations are attracting increasing attention, they still remain a challenging problem due to the complex cross-modal interactions and possible mismatches between training and test data in each modality.

In this section, commonly used types of single-modal representations, such as text and images, are reviewed. These representations often serve as cornerstones for learning multimodal representations. Next, both supervised and unsupervised methods for learning a joint representation space for multiple modalities are introduced. To enable models to handle data samples with missing modalities, the zero-shot learning problem can be solved to increase the similarity of representational spaces across the involved modalities. Finally, inspired by the success of adapting pre-trained LMs to downstream tasks in NLP, methods that leverage large unimodal datasets to improve the learning of multimodal representations are also discussed.

A. Unimodal Embeddings

A distributed representation is a vector that distributes information related to a concept with multiple elements, indicating that elements can be tuned separately to allow more concepts to be encoded efficiently in a relatively low-dimensional space [68]. Such representations can be compared to symbolic representations, such as one-hot encoding, which uses an element with a value of one to indicate the presence of a concept locally and values of zero for other elements. In deep learning, the term "embedding" often refers to a mapping from a one-hot vector representing a word or image category to a distributed representation of real-valued numbers.

1) Visual Representations: Image embeddings can be acquired as output values from the final CNN layers in models that classify images into categories, such as AlexNet [19], VGGNet [20], GoogLeNet [22], and ResNet [21]. AlexNet, GoogLeNet, and ResNet were the winners of the 2012, 2014, and 2015 ImageNet Large Scale Visual Recognition Competition for image classification, respectively [73], [74]. Alternatively, features with more direct relationships to semantics can be used as visual embeddings, such as convolutional features and associated class labels from selected regions identified by object detection models. Models using this approach include the region-based CNN (R-CNN) [23], Fast R-CNN [24], and Faster R-CNN [25]. It should be noted that these models are only a few examples and do not cover all popular CNN structures.

2) Language Representations: Text embeddings can be derived from a neural network language model (NNLM) [69], which estimates the probability of a text sequence by factorizing the sequence into word probabilities using a chain rule for probability. RNN-based NNLMs, such as long short-term memory (LSTM) or gated recurrent unit (GRU) LMs [75], [76], facilitate the use of information from all past words stored in a fixed-length recurrent vector when predicting a current word. In addition to NNLMs, the continuous bag-of-words model, skip-grams, and global vectors (GloVe) are other commonly used methods for word embeddings [70], [77]. A series of deep structured semantic models (DSSMs) have been proposed since 2013 for sentence-level embedding learning based on the optimization of semantic similarity-driven objectives using various neural network structures in pseudo-Siamese network settings [65]–[67], [78]–[81].

Recently, to accomplish downstream natural language understanding tasks with small amounts of training data, many studies have focused on learning general text embeddings by predicting word probabilities using NNLMs with complex structures based on large text corpora. Embeddings from language models [31] use combined embeddings from multiple layers of bidirectional LSTMs for forward and backward propagation. Generative

pre-training [32] and bidirectional encoder representations for transformers (BERT) [33] use the decoder and encoder components of transformer models to estimate the probability of a current subword unit. Besides the word and subword levels, text embeddings can also be learned on the phrase, sentence, and paragraph levels [28], [82].

3) Vector Arithmetic for Word and Image Embeddings: It is well known word embeddings can capture both syntactic and semantic regularities. A famous example showed that the operation vector("King")−vector("Man")+vector("Woman") results in a vector closest to the vector("Queen"), where vector$(\cdot)$ denotes the representation of a word learned by an RNN LM [83]. A similar phenomenon has also been observed for vision embeddings. When using a generative adversarial network (GAN) [84], it has been shown that the operation vector("Man with glasses")−vector("Man")+vector("Woman") results in vector("Woman with glasses") [85], where vector$(\cdot)$ refers to the representation of an image. This result indicates that GANs can capture image representations that disentangle the concept of gender from that of wearing glasses. These findings regarding both text and image representations have encouraged additional studies on joint representations of these two modalities. Additional details regarding GAN-based image generation can be found in Section IV-B.

4) Speaker Representations: Despite speech-related studies are not the focus of this paper, a brief discussion on speaker representation is presented here since it is broadly used in many downstream tasks nowadays [86]–[97]. The i-vector approach estimates a vector for every speaker using factor analysis [98]. Speaker-specific vectors can be jointly trained with the DNN acoustic models [99], [100]. Speaker embeddings can also be derived as the outputs from the penultimate layer of a DNN trained to classify the training set speakers at frame level, namely the d-vectors [101]. Alternatively, the model can be trained to discriminate speakers at utterance level using a statistical pooling layer or a self-attentive layer [102]–[104], and x-vector is the first of such approaches. The training set for x-vectors is often augmented to include different background noises and channels *etc.*, which helps to disentangle the concept of the speaker's voice characteristics from the others. Privacy-preserving is an important issue for speech product in practice, secure binary embeddings can be used to estimate speaker embeddings without exposing the speaker data [105], [106].

B. Multimodal Representations

Although significant progress has been made in terms of learning representations for vision or languages, it is theoretically insufficient to model a complete set of human concepts using only unimodal data. For example, the concept of a "beautiful picture" is grounded in visual representation, so it can be difficult to describe this concept using natural language or other non-visual approaches. Therefore, it is important to learn joint embeddings to leverage the complementarity of multimodal data to represent such concepts more accurately.

1) Unsupervised Training Methods: Joint embeddings for multimodal data can be learned by simply reconstructing raw inputs using multiple streams of deep Boltzmann machines or auto-encoders with shared layers acting as a shared representation space [107]–[109]. Alternatively, based on the development of methods for single modalities, a shared representation space can be constructed by mapping pre-trained representation spaces for the involved individual modalities into a common space. For example, Fang *et al.* proposed a deep multimodal similarity model (DMSM) [110] as an extension of the text modal DSSM to learn embedding representations of text and images in a unified vector space. The simple fusion of word and image embeddings was accomplished using addition or concatenation [109], [111]. The similarity between textual and visual embeddings can be increased through training [112]. In a recent study, the correlation and mutual information between embeddings of different modalities were maximized [113], [114]. Similarly, the distances between word embeddings can be modified according to the similarities between their visual instantiations [115], which are determined by clustering abstract scenes in an unsupervised manner.

Other studies have correlated image regions/fragments with sentence fragments or attribute words to generate fine-grained multimodal embeddings [116] by calculating the alignments between images and sentence fragments automatically. Wu *et al.* unified the embeddings of concepts at different levels, including objects, attributes, relationships, and full scenes [117]. The stacked cross attention network was proposed to learn fine-grained word and image-object aligned embeddings for image-text matching [118]. Additionally, the deep attentional multimodal similarity model (DAMSM) was proposed [51] as an extension of DMSM with attention models to measure the similarity between image sub-regions and words as an additional loss function for text-to-image generation.

2) Supervised Training Methods: Supervised training can be used to improve the learning of multimodal representations. Representations can be factorized into two sets of independent factors: multimodal discriminative factors for supervised training and intra-modality generative factors for unsupervised training [119]. Discriminative factors are shared across all modalities and are useful for discriminative tasks, whereas generative factors can be used to reconstruct missing modalities. Based on detailed text annotations, some researchers have proposed learning word embeddings from their visual co-occurrences (ViCo) when considering natural scene images or image regions [120]. The concept of ViCo is complementary to GloVe text embedding in that it more accurately represents similarities and differences between visual concepts that are difficult to obtain from text corpora alone. Multiple supervised training tasks have been applied to different layers of vision-language encoders [121]. The order of training tasks is determined based on the concept of curriculum learning to increase the complexity of training objectives in a step-by-step manner.

3) Methods for Zero-Shot Learning: Zero-shot learning is often applied to vision-related tasks based on the difficulty of acquiring sufficient labelled images for training for all possible object categories. However, not all types of multimodal representations are suitable for zero-shot learning because certain representations may require pairwise data from different modalities simultaneously. Here, we review methods that rely on additional language sources to overcome this issue. Deep learning-based zero-shot learning begins by developing a

linear mapping layer between different pre-trained embeddings [122], [123]. The deep visual-semantic embedding model was constructed using skip-gram text embeddings and AlexNet visual features. This model allows both types of pre-trained models to be jointly trained through a linear mapping layer [123]. This model was subjected to large-scale test with 1000 known classes and 2000 unknown classes. Better representations could be learned for one-shot and few-shot image retrieval when correlated auto-encoders were used to reconstruct the representations for each modality [124]. A recent work used word labels related and unrelated to a target class to derive visual embeddings from a pre-trained VGG network as positive and negative visual priors, which were used as the inputs for another model to achieve the semantic image segmentation of new object classes that were unseen in the training set [125]. Rich information sources can be used for multiple modalities, including words selected from Wikipedia articles and features derived from multiple CNN layers [126]. Rather than using direct text attribute inputs, sentence embeddings generated by recurrent models can be used as a text interface for zero-shot learning to achieve enhanced results [127].

4) Transformer-Based Methods: Transformers are prevalent sequence-based encoder-decoder models that are formed by stacking many blocks of feedforward layers with multi-head self-attention models, whose parameters are shared temporally [128]. Compared to RNN-based encoder-decoder models [27], such models can provide superior performance on long sequences based on the removal of the first-order Markovian assumption imposed on RNNs. BERT, which is the encoder component of a transformer pre-trained on a large text corpus as a masked LM, is a standard choice for text embeddings for downstream tasks. Therefore, it is natural to generalize text-only BERT to cover images and derive pre-trained bimodal embeddings.

A straightforward method for extending unimodal BERT to bimodal applications is to include new tokens to indicate visual feature inputs, such as those proposed in [129]–[133]. Additionally, the transformer model can be modified by introducing an extra encoder or attention structures for visual features [134]–[136]. Additional details regarding modified structures can be found in Section III-B. Furthermore, a recent NLP study suggested that multitask learning can improve the generalization ability of BERT representations [137]. Therefore, most of the aforementioned bimodal BERT-based models adopt multitask training to improve their performance on downstream tasks, such as VQA, and image and video captioning *etc*.

III. Fusion

Fusion is a key research topic in multimodal studies, which integrates information extracted from different unimodal data sources into a single compact multimodal representation. There is a clear connection between fusion and multimodal representations. We classify an approach into the fusion category if it focuses on architectures for integrating unimodal representations for a particular task. Fusion methods can be divided based on the stage in which fusion occurs during the associated procedures. Because early and late fusion can suppress either intra- or inter-modality interactions, recent studies have focused on intermediate methods that allow fusion to occur on multiple layers of a deep model.

This section presents a review of intermediate fusion not only because it is more flexible but also because the boundaries between stages are less clear based on the use of unimodal features derived from pre-trained backbone models. Three types of methods that are mainly used to fuse text with image features are considered, namely, simple operation-based, attention-based, and tensor-based methods.

A. Simple Operation-Based Fusion

In deep learning, vectorized features from different information sources can be integrated using simple operations, such as concatenation or weighted sums, which often have few or no associated parameters because the joint training of deep models can adapt layers for high-level feature extraction to adjust to the required operations.

- Concatenation can be used to combine either low-level input features [138]–[140] or high-level features extracted by pre-trained models [140]–[142].
- For weighted sums with scalar weights, an iterative method that requires the pre-trained vector representations to have the same number of elements arranged in an order that is suitable for element-wise addition has been proposed [143]. This can be achieved by training a fully connected layer for dimension control and reordering for each modality.

A recent study [144] employed neural architecture search with progressive exploration [145]–[147] to find suitable settings for a number of fusion functions. Each fusion function was configured in terms of which layers to fuse and whether to use concatenation or a weighted sum as the fusion operation.

B. Attention-Based Fusion

Attention mechanisms are widely used for fusion. Attention mechanisms often refer to the weighted sum of a set of vectors with scalar weights that are dynamically generated by a small "attention" model at each time-step [148], [149]. Multiple glimpses (output heads) are often used to generate multiple sets of dynamic weights for summation, which can preserve additional information by concatenating the results derived from each glimpse. When applying attention mechanisms to an image, image feature vectors that are relevant to different regions are weighted differently to produce an attended image vector.

1) Image Attention: An LSTM model for text question processing was extended by incorporating an image attention model conditioned on previous LSTM hidden states, whose inputs were concatenations of the current word embedding and attended image feature [150]. The final LSTM hidden state was used as a fused multimodal representation to predict an answer for pointing and grounded VQA. The attention model for an RNN-based encoder-decoder model was used to assign attention weights to image features for image captioning [151]. Additionally, for VQA, an attention model conditioned on both images and query feature vectors was applied to pinpoint image regions

that were relevant to the answer [152]. Similarly, stacked attention networks (SANs) have been proposed to use multiple layers of attention models to query an image multiple times to infer an answer progressively and simulate a multi-step reasoning procedure [153]. In each layer, a refined query vector is generated and sent to the next layer by adding the previous query vector to the attended image vector produced using the current attention model. A spatial memory network is a multi-hop method for VQA that aligns words to image regions in a first hop and assigns image attention based on entire questions in a second hop to derive an answer [154].

A dynamic memory network is augmented to use separate input modules to encode questions and images. This type of network uses attention-based GRUs to update episodic memory iteratively and retrieve required information [155]. The bottom-up and top-down attention method (Up-Down), as its name suggests, simulates the human visual system using a combination of two visual attention mechanisms [156]. The bottom-up attention mechanism proposes a set of salient image regions identified by a Faster R-CNN and the top-down attention mechanism performs concatenation of visual and linguistic features to estimate attention weights and produce an attended image feature vector for image captioning or VQA. The attended image feature vector can be fused with linguistic features again by computing an element-wise product. Complementary image features derived from different models, such as ResNet and Faster R-CNN, can be used for multiple image attention mechanisms [157]. Furthermore, the inverse of image attention, which generates attended text features from image and text inputs, can be used for text-to-image generation [51], [158].

2) Symmetric Attention for Images and Text: In contrast to the aforementioned image attention mechanisms, co-attention mechanisms use symmetric attention structures to generate not only attended image feature vectors but also attended language vectors [159]. Parallel co-attention uses a joint representation to derive image and language attention distributions simultaneously. In contrast, alternating co-attention has a cascaded structure that first generates an attended image vector using linguistic features and then generates an attended language vector using the attended image vector.

Similar to parallel co-attention, a dual attention network (DAN) estimates attention distributions for images and languages simultaneously to derive attended feature vectors [160]. Such attention models are conditioned on both features and memory vectors related to relevant modalities. This is a key difference compared to co-attention because memory vectors can be iteratively updated at each reasoning step using repeated DAN structures. Memory vectors can be either shared for VQA or modality-specific for image-text matching. Stacked latent attention (SLA) improves SANs by concatenating original attended image vectors with values from earlier layers in the attention model to retain latent information from intermediate reasoning stages [161]. A parallel co-attention like twin-stream structure is also included to assign attention to both image and language features, which facilitates iterative reasoning using multiple SLA layers. Dual recurrent attention units implement a parallel co-attention structure using LSTM models for text and images to assign attention weights to each input location in representations obtained by convoluting image features using a stack of CNN layers [162]. To model high-order interactions between modalities, the high-order correlations between two data modalities can be computed as the inner product of two feature vectors and used to derive attended feature vectors for both modalities [163].

3) Attention in a Bimodal Transformer: As mentioned in Section II-B4, the bimodal extensions of BERT rely on different tokens to indicate whether a vector is a word or image fragment. Attention models then fuse images with words in bimodal input sequences [129]–[133]. OmniNet uses a gated multi-head attention model in each decoder block to fuse vectors from other modalities with those produced for the current modality by the previous layers in each block [136]. LXMERT uses independent encoders to learn intra-modality features for each modality and a cross-modality encoder on a higher level to learn cross-modality features using additional cross-attention layers [134]. ViLBERT extends BERT to include two encoder streams to process visual and textual inputs separately. These features can then interact through parallel co-attention layers [134].

4) Other Attention-Like Mechanisms: The gated multimodal unit is a method that can be viewed as assigning attention weights to images and text based on gating [164]. This method computes a weighted sum of visual and textual feature vectors based on dimension-specific scalar weights generated dynamically by a gating mechanism. Similarly, element-wise multiplication can be used to fuse visual and textual representations. These fused representations are then used to create the building blocks for a multimodal residual network based on deep residual learning [165]. A dynamic parameter prediction network uses a dynamic weight matrix to transform visual feature vectors, whose parameters are dynamically generated by hashing text feature vectors [166].

C. Bilinear Pooling-Based Fusion

Bilinear pooling is a method that is often used to fuse visual feature vectors with textual feature vectors to create a joint representation space by computing their outer product, which facilitates multiplicative interactions between all elements in both vectors. This method is also referred to as second-order pooling [167]. In contrast to simple vector combination operations (assuming each vector has n elements), such as a weighted sum, element-wise multiplication, or concatenation, which result in n- or $2n$-dimensional representations, bilinear pooling generates an n^2-dimensional representation by linearizing the matrix generated by the outer product into a vector, meaning this method is more expressive. Bilinear representations are often linearly transformed into output vectors using a two-dimensional weight matrix, which is equivalent to using a three-dimensional tensor operator to fuse two input feature vectors. Each feature vector can be extended with an extra value of one to preserve single-modal input features in the bilinear representation when calculating an outer product [168]. However, based on its high dimensionality (typically on the order of hundreds of thousands to a few million dimensions), bilinear pooling often requires the decomposition of weight tensors to allow the associated model to be trained properly and efficiently.

1) Factorization for Bilinear Pooling: Because bilinear representations are closely related to polynomial kernels, various low-dimensional approximations can be used to acquire compact bilinear representations [169]. Count sketches and convolutions can be used to approximate polynomial kernels [170], [171], leading to multimodal compact bilinear pooling (MCB) [172]. Alternatively, by imposing low ranks on weight tensors, multimodal low-rank bilinear pooling (MLB) factorizes three-dimensional weight tensors for bilinear pooling into three two-dimensional weight matrices [173]. Specifically, visual and textual feature vectors are linearly projected onto low-dimensional modality-specific factors by two input factor matrices. These factors are then fused using element-wise multiplication, followed by linear projection using the third matrix for output factors. Multimodal factorized bilinear pooling (MFB) modifies MLB with an extra operation to pool element-wise multiplication results by summing the values within each non-overlapping one-dimensional window [174]. Multiple MFB models can be cascaded to model high-order interactions between input features, which is referred to as multi-modal factorized high-order pooling (MFH) [175].

MUTAN, which is a multimodal tensor-based Tucker decomposition method, uses Tucker decomposition [176] to factorize the original three-dimensional weight tensor operator into a low-dimensional core tensor and the three two-dimensional weight matrices used by MLB [177]. Core tensors model the interactions across modalities. MCB can be considered as MUTAN with fixed diagonal input factor matrices and a sparse fixed core tensor, while MLB can be considered as MUTAN with the core tensor set to the identity tensor. Recently, BLOCK, which is a block-based super-diagonal fusion framework, was proposed to perform block-term decomposition [178] to compute bilinear pooling [179]. BLOCK generalizes MUTAN as a summation of multiple MUTAN models to provide richer modeling of the interactions between modalities. MUTAN core tensors can be arranged as super-diagonal tensors, similar to the submatrices of a block diagonal matrix. Furthermore, bilinear pooling can be generalized to more than two modalities, such as using outer products to model the interactions among representations for video, audio, and language [168], [180].

2) Bilinear Pooling and Attention Mechanisms: Bilinear pooling can be combined with attention mechanisms. MCB/MLB fused bimodal representations can be used as input features for an attention model to derive an attended image feature vector, which is then fused with a textual feature vector by using MCB/MLB again to form a final joint representation [172], [173]. MFB/MFH can be used for alternating co-attention to learn joint representations [174], [175]. A bilinear attention network (BAN) uses MLB to fuse images and text to produce a bilinear attention map representing an attention distribution, which is then used as a weight tensor for bilinear pooling to fuse images and text features again [181].

IV. Applications

This section discusses selected applications for multimodal intelligence that combine vision and language, including image captioning, text-to-image generation, and VQA. Note that there are other common applications, including text-based image retrieval [118] and visual-and-language navigation [182], that we have not included in this review owing to space limitations.

- Image captioning is a task that aims to generate a natural language description of an image automatically. It requires a level of image understanding beyond that provided by typical image recognition and object detection methods.
- The inverse of image captioning is text-to-image generation, which generates image pixels according to a description or keywords provided by humans.
- VQA is related to image captioning. It often takes an image and a free-form, open-ended natural language question about the image as inputs and then outputs a classification result as an answer. Natural language understanding is necessary because questions are free in form. Other capabilities, such as knowledge-based reasoning and commonsense reasoning, are also important because questions are open-ended.
- Visual reasoning can be included in all of the aforementioned tasks. However, only methods related to VQA are reviewed here.

Detailed task specifications, datasets, and selected work for each task will be introduced in this section.

A. Image Captioning

Image captioning is a task that requires the generation of a text description of an image [183]. It is one of the first tasks involving the multimodal combination of images and text. We mainly review deep learning-based methods. The image captioning task can be divided into several sub-tasks, allowing captions to be generated in a step-by-step manner [40], [110], [184]. For example, a deep CNN model can be trained to detect words in images, and then a log-linear language model can be used to compose words into sentences [110]. Similarly, image features can be fed into a log-linear language model to generate sentences [184]. In contrast, exact matching of objects in images and words in sentences attempts to determine if an image and sentence match with each other [40].

Similar to RNN-based encoder-decoder methods for machine translation [27], another approach was proposed to generate captions from images in an end-to-end manner using an encoder-decoder architecture [185]–[187]. In this type of model, a CNN, which is typically pre-trained using ImageNet [73], encodes an image into a continuous vector, which is then fed into an RNN/LSTM decoder to generate captions directly. These types of methods all use the same basic architecture, but they vary slightly in their choices of CNN parameters and how image vectors are fed into decoders. Although this method is powerful and convenient, the encoder-decoder architecture lacks the ability to capture the fine-grained relationships between objects in images and words in sentences. To overcome this issue, the attention-based encoder-decoder model has been proposed and has become the standard benchmark for this task [188]. In the attention encoder-decoder model, prior to generating the next word, the decoder first calculates matching scores (attention) with objects in an image and then considers the weighted image features to generate the next token. There have been many

studies that have attempted to improve the attention model by incorporating additional structures. For example, Lu *et al.* added a gate at every decoding step to determine if the next word should be generated using image information [189]. Additionally, detected words and image features can be combined as inputs for the decoder network [45], [190]. More recently, many studies have incorporated extra structures/knowledge from either images [156] or text [191]. Specifically, an object detector was used to localize features for an image object and generate captions based on the localized features [156]. This method improved upon the previous state-of-the-art model by a wide margin in terms of a variety of evaluation metrics.

Image captions with rich information can be generated by incorporating external knowledge. For example, based on a database of celebrities [192], a the CaptionBot application was developed to describe the components (such as activities) of an image, as well as who is related to each component if the people in the image can be recognized [193]. In addition to generating factual descriptions of images, other approaches have been proposed for explicitly controlling the style [194], semantic content [190], and diversity [195] of generated captions.

B. Text-to-Image Generation

Text-to-image generation, which relies on natural language to control image generation, is a fundamental problem in computer vision. It is considered to be a difficult problem because it involves at least two tasks: high-quality image generation and language understanding. Generated images must be both visually realistic and semantically consistent with language descriptions. Deep learning-based text-to-image generation can be dated back to the use of LSTM for iterative handwriting generation [196]. This iterative image generation method was later extended to create the deep recurrent attentive writer (DRAW) method, which combines an LSTM-based sequential variational auto-encoder (VAE) with a spatial attention mechanism [197]. The alignDRAW method modifies DRAW to use natural language-based descriptions to synthesize images with general content [198]. An attention model is used to compute the alignment between input words and iteratively drawn patches. GAN-based methods have become the major focus of more recent text-to-image generation studies, potentially because the discriminators of GANs can serve as reasonable criteria for evaluating synthesized images, which is difficult to achieve using other methods. The following subsection provides an overview of some GAN-based methods, including the basic settings and solutions for some important problems, such as the generation of high-quality images, semantic consistency between images and text, and the layout control of images *etc*.

1) GAN-Based Methods: Compared to VAE, conditional-GAN (CGAN) can synthesize more compelling images of specific categories that a human may even mistake for real images [199], [200]. A GAN model consists of a generator that synthesizes candidates based on input noise and a discriminator that evaluates the candidates. Adversarial training is employed to train the generator to capture true data distributions so the discriminator can no longer discriminate between synthesized data and real data [84]. CGAN extends the standard GAN structure by generating additional category labels for both the generator and discriminator. The GAN-INT-CLS method facilitates the synthesis of visually plausible 64×64 images by using embeddings of natural language descriptions to replace category labels in CGAN [201]. Automatic evaluation of the quality of text-conditioned images can be less straightforward. To determine the discriminability of GAN-generated images, the inception score [202] and Fréchet inception distance [203] metrics are often used. Multi-scale structural similarity [204] is used to evaluate the diversity of images. To evaluate whether a generated image is semantically consistent with an input text description, R-precision [51] and visual-semantic similarity [205] are commonly used metrics.

2) Generating High-Quality Images: Although they basically reflect the meanings of descriptions, it has been found that the images produced by GAN-INT-CLS do not contain fine-grained details or vivid objects. This shortcoming motivated the development of the StackGAN method [206]. StackGAN decomposes image synthesis into more manageable sub-problems through a sketch-refinement process by stacking two separately trained CGANs. The first GAN produces 64×64 low-resolution images by sketching the primitive shapes and colors of objects based on text. The second GAN is then trained to generate 256×256 images by rectifying defects and adding compelling details to the low-resolution images generated by the first GAN. StackGAN++ improves upon this idea by incorporating an additional GAN to generate 128×128 images between the two GANs discussed above and training all GANs jointly [207]. To ensure the generated images semantically match the text precisely, the attentional GAN (AttnGAN) was proposed. This method also stacks three GANs targeting different image resolutions [51]. The first GAN is trained on sentence embeddings, and the next two GANs are trained on bimodal embeddings produced by attention models fusing word-level features with low-resolution images. It has been shown that attention mechanisms can help GANs focus on the words that are the most relevant to the sub-regions drawn in each stage. In addition to stacking generators, it has been shown that high-resolution images can be generated using dynamic memory modules [208]. The progressive growth of a GAN begins with training a one-layer generator and one-layer discriminator to synthesize 4×4 images. This method then progressively adds more layers to both models to increase image resolution up to 1024×1024 [209].

3) Generating Semantically Consistent Images: To improve the semantic consistency between relevant images and text features, DAMSM was proposed for AttnGAN [51]. The hierarchically-nested discriminator GAN (HDGAN) [205] handles the same problem by leveraging hierarchical representations with additional adversarial constraints to discriminate not only real/fake image pairs but also real/fake image-text pairs at multiple image resolutions in the discriminator. Similarly, the text-conditioned auxiliary classifier GAN (TAC-GAN) introduces an additional image classification task into the discriminator [210], while the text-conditioned semantic classifier GAN (Text-SeGAN) trains classifiers by using regression tasks to estimate the semantic relevance between images and text [211]. As an analogue to cycle consistency [212], MirrorGAN

was proposed to improve the semantic consistency between two modalities using an additional image captioning module [213].

4) Semantic Layout Control for Complex Scenes: Despite the success in the generation of realistic and semantically consistent images for single objects, such as birds [214] or flowers [215], state-of-the-art text-to-image generation methods still struggle to generate complex scenes containing many objects and relationships, such as those in the Microsoft COCO dataset [216]. In the pioneering work in [217], both text descriptions and the locations of objects specified by keypoints or bounding boxes were used as inputs. Later, detailed semantic layouts, such as scene graphs, were used to replace natural language sentences with more direct descriptions of objects and their relationships [218], [219]. Additionally, efforts have been made to maintain natural language inputs while incorporating the concept of semantic layouts. Hinz *et al.* included extra object pathways in both a generator and discriminator to control object locations explicitly [220]. Hong *et al.* employed a two-stage procedure that first constructs a semantic layout automatically from an input sentence using LSTM-based box and shape generators and then synthesizes images using image generators and discriminators [221]. Because fine-grained word/object-level information is not explicitly used for generation, such synthesized images do not contain sufficient details to make them look realistic. The object-driven attentive GAN (Obj-GAN) improves upon the two-stage generation concept using a combination of an object-driven attentive image generator and object-wise discriminator [158]. At every generation step, the generator uses a text description as a semantic layout and synthesizes image regions within a bounding box by focusing on the words that are the most relevant to the object within that box. Obj-GAN is more robust and interpretable compared to other GAN methods and significantly improves object generation quality for complex scenes.

5) Additional Topics: In addition to layouts, other types of fine-grained control for image generation have been discussed in the literature. Attribute2Image [222] uses various attributes for face generation, such as age and gender. This concept has also been adapted to face editing to remove beards or change hair colors [223]. The text-adaptive GAN [224] facilitates the semantic modification of input images of birds and flowers based on natural language. Lao *et al.* proposed to enforce the learning of representation content and styles as two disentangled variables by using a dual inference mechanism based on cycle-consistency for text-to-image generation [225]. The success of these methods demonstrates that GANs are able to learn some semantic concepts as disentangled representations, as discussed in Section II-A3. Text2Scene is another noteworthy method that generates compositional scene representations from natural language in a step-by-step manner without using GANs [226]. It has been shown that with minor modifications, Text2Scene can generate cartoon-like, semantic layout, and real image-like scenes. Dialogue-based interactions have also been studied to control image synthesis by improving complex scene generation progressively [227]–[231]. Text-to-image generation has also been extended to multiple images or videos, where visual consistency among generated images is required [232]–[234].

C. Visual Question Answering

1) Task Definition: VQA extends text-based QA from NLP by asking questions related to visual information presented in an image or video clip. Image-based VQA is often considered as a visual Turing test, where a system is required to understand any form of natural language-based questions and answer them in a natural manner. However, it is often simplified as a classification task defined in different ways to focus on different core problems [46], [47], [150], [235], [236]. Initial works generated questions using templates or by converting descriptive sentences using syntax trees [235], [237]. Later studies focused on the use of free-form natural language questions authored by either humans or powerful deep generative models, such as GANs and VAEs [47], [237]–[239]. In contrast to open-ended questions, which are presented in complete sentence form, possible answers are often presented as a large set of classes (*e.g.*, 3000 classes) related to yes/no answers, counts, object classes, and instances *etc*. To focus on core understanding and reasoning problems, VQA can be simplified to classify visual and textual features into answer-related classes.

Alternatively, VQA can be defined to select outputs among multiple (e.g., four) choices, where each choice is associated with an answer presented in the form of a natural language sentence [150]. This setup can be implemented as a classification problem based on the features of images, questions, and answer candidates [172]. There are other types of VQA task definitions as well, such as the Visual Madlibs dataset, which requires answering questions using a "fill-in-the-blanks" system [48]. Furthermore, visual dialogue can be viewed as the answers to a series of questions grounded in images [240], [241]. This method extends VQA by requiring the generation of more human-like responses and the inference of context based on dialogue history.

2) Common Datasets and Approaches: The first VQA data set, which is called DAQUAR, uses real-world images combined with both template-based and human-annotated questions [235]. COCO-QA contains more QA pairs than DAQUAR because it converts image descriptions from the MS COCO dataset into questions [237]. Such questions are generally easier to answer because they allow models to rely more on rough images, rather than logical reasoning. VQA v1 and v2 are the most popular datasets for VQA. These datasets consist of open-ended questions with both real and abstract scenes [47], [242]. A VQA challenge based on these datasets has been held annually as a workshop since 2016. Visual7W is a portion of the Visual Genome dataset for VQA containing multiple choices [150]. It contains questions related to the concept of "what", "who", and "how" for spatial reasoning and "where", "when", and "why" for high-level commonsense reasoning. The seventh type of questions in Visual7W are "which" questions, which are also referred to as pointing questions. The answer choices for these questions are associated with the bounding boxes of objects in images. Approaches designed for these datasets often focus on fusing image and question vectors with the aforementioned discussed attention- and bilinear-pooling-based methods, including SAN, co-attention, Up-Down, MCB, MLB, and BAN *etc*.

3) Integrating External Knowledge Sources: Since most of the VQA questions in the aforementioned datasets focus on simple counting, color, and object detection problems that do not require any external knowledge, a possible extension of these tasks is to include more difficult questions that require knowledge beyond what the questions entail or what information is contained in images. Both knowledge-based reasoning for VQA and fact-based VQA datasets incorporate structured knowledge bases, which often require additional steps to query knowledge bases, meaning the corresponding methods are no longer trainable in an end-to-end manner [243], [244]. In contrast to structured knowledge bases, outside-knowledge VQA uses external knowledge in the form of natural language sentences collected by retrieving Wikipedia articles using search queries extracted from questions. Additionally, an ArticleNet model is trained to find answers in retrieved articles [245].

4) Discounting Language Priors: Although significant achievements have been made, recent studies have pointed out that common VQA benchmarks suffer from strong and prevalent priors (*e.g.*, "most bananas are yellow" and "the sky is mostly blue"), which can often cause VQA models to overfit statistical biases and tendencies in answer distributions. This issue largely circumvents the need to understand visual scenes. Based on the objects, attributes, and relationships provided by the scene graphs of Visual Genome, a new dataset called GQA was created to reduce biases by generating questions using a functional program that controls reasoning steps [246]. New splits for VQA v1 and VQA v2 were generated to provide different answer distributions for every question in the training and test sets. These splits are referred to as VQA under challenging priors (VQA-CP v1 and VQA-CP v2) [247]. Other methods have been proposed to handle biased priors using adversarial training or additional training-only structures [248], [249].

5) Additional Issues: Another problem that current VQA methods suffer from is low robustness against linguistic variation in questions. A dataset called VQA-Rephrasings modifies the VQA v2 validation set with human-authored rephrasing of questions [212]. Additionally, a cycle-consistency-based [250] method was proposed to improve linguistic robustness by enforcing consistency between original and rephrased questions, as well as between true answers and answers predicted based on original and rephrased questions. Zhang *et al.* suggested that attention mechanisms can cause VQA models to suffer from counting object proposals and an additional model component was proposed as a solution [251]. Additionally, it is known that current VQA methods cannot read text from images. A method was proposed to address this problem by fusing text extracted from images using optical character recognition [252]. VizWiz is a goal oriented VQA dataset collected by blind people capturing potentially low-quality images and asking questions in spoken English. This dataset includes many text-related questions [253]. To learn rare concepts that humans may talk more likely than the commonsense knowledge, active learning, which allows a model to seek labels selectively for more informative examples, has been applied to VQA to reduce data annotation efforts [254]–[256].

D. Visual Reasoning

This section focuses on the study of a very intriguing problem called visual reasoning, which focuses on how to accomplish accurate, explicit, and expressive understanding and reasoning. Visual reasoning is related to many language- and vision-based bimodal tasks, such as captioning and text-to-image generation. However, in this section, we mostly focus on methods related to VQA because visual reasoning is particularly important when answering complicated questions. SANs are often considered as being closely related to implicit visual reasoning because their stacked structures can be viewed as performing multiple reasoning steps. Feature-wise linear modulation was proposed to refine visual features iteratively using feature-wise affine transformations based on scaling factors and bias values generated dynamically from textual features [257]. Multimodal relational networks (MuRel) also have structures with multiple MuRel cells based on bilinear pooling, which can be used iteratively [258].

1) Neural Module Network-Based Methods: A neural module network (NMN) consists of a collection of jointly trained neural "modules" combined into a deep model for answering questions [259]. A dependency parser first helps convert natural language questions into a fixed and rule-based network format and specifies both the set of modules used to answer questions and the connections between modules. Next, a deep model is assembled based on the target question format to generate answer predictions. SHAPES, which is a synthetic dataset consisting of complex questions regarding simple arrangements of ordered shapes, was proposed to focus on the compositional aspects of questions [259]. A later study trained a model layout predictor jointly with module parameters by re-ranking a list of layout candidates using reinforcement learning. This method is called a dynamic NMN [260]. Modules for "find" or "relate" operations use attention models to focus on one or two regions in an input image and make the execution of assembled deep models similar to running a functional program [260]. An end-to-end version of the NMN used an RNN question encoder to convert input questions into layout policies without requiring the aid of a parser [261]. This work was based on a relatively new dataset called compositional language and elementary visual reasoning diagnostics (CLEVR). As its name suggests, CLEVR is a synthetic diagnostic dataset for testing a range of visual reasoning abilities related to objects and relationships with minimal biases and detailed annotations describing the type of reasoning each question requires [262]. Other implementations of the NMN include the program generator and execution engine method (PG+EE), which shares generic designs among certain operations [263]; the stack-NMN, which improves the parser and incorporates question features into modules [264]; and the transparency-by-design network, which redesigns some modules from PG+EE to maintain the transparency of the reasoning procedure [265].

2) Other Types of end-to-end Reasoning Methods: Another end-to-end approach is the memory, attention, and composition (MAC) network, which decomposes questions into a series of attended reasoning steps and performs each step using a

recurrent MAC cell that maintains a separation between control and memory hidden states. Each hidden state is generated by an ANN model constructed based on attention and gating mechanisms [266]. Recently, both deterministic symbolic programs and probabilistic symbolic models have been used as execution engines for generated programs to improve transparency and data efficiency, resulting in the creation of the neural-symbolic VQA (NS-VQA) and probabilistic neural-symbolic models, respectively [267], [268]. As an extension of the NS-VQA, the neuro-symbolic concept learner (NS-CL) uses a neuro-symbolic reasoning module to execute programs based on scene representations. The NS-CL can have its program generator, reasoning module, and visual perception components trained jointly in an end-to-end manner without requiring any component-level supervision [269]. Its perception module learns visual concepts based on language descriptions of objects and facilitates learning new words and parsing new sentences.

We conclude this section by reviewing the relationship network (RN), which has a simple structure that uses an ANN as a function to model the relationships between any pair of visual and textual features. The resulting output values are then accumulated and transformed by another ANN [270]. Although the RN simply models relationships without any form of inductive reasoning, it achieves very high VQA accuracy on the CLEVR dataset. This inspires a re-thinking of the connections between correlation and induction.

V. Summary and Prospects

This paper reviewed the topics of modeling and machine learning across multiple modalities based on deep learning with a focus on the combination of vision and natural language. We organized many different works from the language-vision multimodal intelligence field according to three factors: multimodal representations, fusion of multimodal signals, and applications of multimodal intelligence. In the section on representations, both single-modal and multimodal representations were reviewed based on the key concept of embedding. Multimodal representations unify relevant signals from different modalities into the same vector space for general downstream tasks. For multimodal fusion, special architectures, such as attention mechanisms and bilinear pooling, were discussed. In the application section, three selected areas of broad interest were presented: image captioning, text-to-image generation, and VQA. A set of visual reasoning methods for VQA was also discussed. Our review covered task definition, dataset specification, development of commonly used methods, as well as issues and trends. We hope this review will promote future studies in the emerging field of multimodal intelligence.

In the future, in addition to the aforementioned research topics, we also want to highlight the following three directions.

A. Multimodal Knowledge Learning

Multiple knowledge bases related to multimodal datasets have been constructed in recent years, such as MS-Celeb-1M [271], which benchmarks recognition of one million celebrities in images and links them to their corresponding information in freebase [272]. In this area, the automatic acquisition of commonsense knowledge from multimodal data can be expected in the near future. Massive amounts of information, including entities, actions, attributes, concepts, and relationships, can be learned from massive amounts of image and video data to construct models covering broad and structured commonsense knowledge. Such models will provide great value for applications related to commonsense reasoning. However, problems that need to be resolved to accomplish this goal include the following:

- Defining commonsense;
- Constructing multimodal datasets and learning commonsense knowledge from them efficiently and effectively;
- Determining which tasks to work on to verify the capabilities of novel algorithms while demonstrating the importance of commonsense;
- Updating previously learned commonsense knowledge.

B. Multimodal Emotional Intelligence

Advanced emotional intelligence is a cognitive ability unique to humans. Communication between humans involves rich emotions and multiple modalities. To construct a highly anthropomorphic human-computer interaction agent, machines must understand and generate multimodal emotional content and empathize with humans. Fundamental research in this area can not only help us to understand the mechanisms of cognitive intelligence, but also has great value for many real-world applications. However, the difficulties of multimodal emotional intelligence include the following:

- Perceiving and aligning the subtle expression of emotions in different modalities;
- Ensuring the consistency and rationality of data across all modalities [273];
- Acquiring the core representations and intensities of emotions that are potentially modality-invariant [274] .

C. Large-Scale Complex Goal-Oriented Multimodal Intelligent Human-Computer Interaction System

The intelligentization of service industries is both a large opportunity and large technical challenge for artificial intelligence. Considering e-commerce as an example, this field faces challenges related to ultra-large-scale data and complex human-computer interactions in the full retail chain. These problems require large-scale, complex, and task-oriented multimodal intelligent human-computer interaction technologies to serve hundreds of millions of users in a personalized and highly efficient way. To this end, opportunities exist in terms of promoting open-source and open-license frameworks for multimodal human-computer interaction systems, constructing large-scale datasets and algorithm verification platforms, and conducting fundamental research on multimodal intelligence. Breakthroughs related to techniques in these areas will also promote the intelligentization of broader service industries.

Regarding the goal of constructing an agent that can perceive multimodal information and use the connections between different modalities to improve its cognitive ability, research on multimodal intelligence is still in its infancy. However, it

has already achieved significant progress and become a very important branch of the development of artificial intelligence.

ACKNOWLEDGMENT

The authors are grateful to the editor and anonymous reviewers for their valuable suggestions that helped to make this paper better.

REFERENCES

[1] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Sci.*, vol. 313, pp. 504–507, 2006.
[2] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, pp. 1–127, 2009.
[3] L. Deng and Y. Dong, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, pp. 197–387, 2014.
[4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.
[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learn.* Cambridge, MA, USA: MIT Press, 2016.
[7] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," presented at the Int. Conf. Neural Inf. Process. Syst. Workshop, Vancouver, Canada, 2010.
[8] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep autoencoder," presented at the Interspeech, Makuhari, Japan, 2010.
[9] L. Deng, "An overview of deep-structured learning for information processing," presented at the Asian-Pacific Signal Inf. Annu. Summit Conf., Xian, China, 2011.
[10] D. Yu, L. Deng, F. Seide, and G. Li, "Discriminative pre-training of deep nerual networks," U.S. Patent 9,235,799, 2011.
[11] G. Dahl, D. Yu, and L. Deng, "Large-vocabulry continuous speech recognition with context-dependent DBN-HMMs," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, 2011, pp. 4688–4691.
[12] L. Deng *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Vancouver, Canada, 2013, pp. 8604–8608.
[13] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[14] F. Seide, L. Gang, and Y. Dong, "Conversational speech transcription using context-dependent deep neural networks," presented at the Interspeech, Florence, Italy, 2011.
[15] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Magaz.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[16] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8599–8603.
[17] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Berlin, Germany: Springer, 2015.
[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998.
[19] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," presented at the Int. Conf. Neural Inf. Process. Syst., Lake Tahoe, 2012.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," presented at the Int. Conf. Learn. Representations, San Diego, 2015.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Las Vegas, 2016, pp. 770–778.
[22] C. Szegedy *et al.*, "Going deeper with convolutions," presented at the Conf. Comput. Vision Pattern Recognit., Boston, 2015.
[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 580–587.
[24] R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vision*, Araucano Park, 2015, pp. 1440–1448.
[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," presented at the Int. Conf. Neural Inf. Process. Syst., Montreal, Canada, 2015.
[26] G. Mesnil *et al.*, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 3, pp. 530–539, Mar. 2015.
[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," presented at the Int. Conf. Learn. Representations, San Diego, CA, USA, 2015.
[28] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," presented at the Int. Conf. Neural Inf. Process. Syst., Montreal, Canada, 2014.
[29] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.
[30] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," presented at the Conf. Empir. Methods Natural Lang. Process., Lisbon, Australia, 2015.
[31] M. Peters *et al.*, "Deep contextualized word representations," presented at the Conf. North Amer. Chapter Assoc. Comput. Linguistics, New Orleans, LA, USA, 2018.
[32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/language understanding paper.pdf
[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," presented at the Conf. North Amer. Chapter Assoc. Comput. Linguistics-Human Lang. Technologies, Minneapolis, MN, USA, 2019.
[34] H.-Y. Shum, X. He, and D. Li, "From Eliza to Xiaoice: Challenges and opportunities with social chatbots," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, pp. 10–19, 2018.
[35] S. Bengio, L. Deng, L. Morency, and B. Schuller, "Perspectives on predictive power of multimodal deep learning: Surprises and future directions," in *The Handbook of Multimodal-Multisensor Interfaces*. San Rafael, CA, USA: Morgan & Claypool, ch. 14, 2019.
[36] L. Deng and Y. Liu, *Deep Learning in Natural Language Processing*. Berlin, Germany: Springer, 2018.
[37] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, 2014, pp. 787–798.
[38] L. Yu, P. Poirson, S. Yang, A. Berg, and T. Berg, "Modeling context in referring expressions," presented at the Eur. Conf. Comput. Vision, Amsterdam, The Netherlands, 2016.
[39] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and L. S., "Flickr30k entities: Collecting region-to phrase correspondences for richer image-to-sentence models," presented at the Int. Conf. Comput. Vision, Araucano Park, Chile, 2015.
[40] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," presented at the Conf. Comput. Vision Pattern Recognit., Boston, MA, USA, 2015.
[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," presented at the Conf. Comput. Vision Pattern Recognit., Boston, MA, USA, 2015.
[42] J. Johnson, A. Karpathy, and F.-F. Li, "Densecap: Fully convolutional localization networks for dense captioning," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.
[43] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.
[44] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.
[45] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4651–4659.
[46] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," presented at the Nat. Acad. Sci., 2015.
[47] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. Int. Conf. Comput. Vision*, Araucano Park, Chile, 2015, pp. 2425–2433.
[48] L. Yu, E. Park, A. Berg, and T. Berg, "Visual madlibs: Fill in the blank description generation and question answering," in *Proc. Int. Conf. Comput. Vision*, Boston, MA, USA, 2015, pp. 2461–2469.
[49] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," presented at the Eur. Conf. Comput. Vision, Amsterdam, The Netherlands, 2016.

[50] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," presented at the Int. Conf. Mach. Learn., New York, NY, USA, 2016.

[51] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.

[52] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.

[53] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[54] M. Cookea, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, 2006.

[55] T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Machi. Intell.*, to be published.

[56] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for video broadcast news: Some fusion techniques," in *Proc. IEEE 3rd Workshop Multimedia Signal Process.*, Copenhagen, Denmark, 1999, pp. 161–167.

[57] Z. Wu, L. Cai, and H. Meng, "Multi-level fusion of audio and visual features for speaker identification," in *Advances in Biometrics* D. Zhang and A. Jain, eds., Berlin, Germany: Springer, 2005, pp. 493–499.

[58] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[59] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, May 2018.

[60] J. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," presented at the Interspeech, Graz, Austria, 2019.

[61] J. Wu *et al.*, "Time domain audio visual speech separation," in *Proc. Autom. Speech Recognit. Understanding Workshop*, Singapore, 2019, pp. 667–673.

[62] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, pp. 112, 1–11, 2018.

[63] T. Afouras, J. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," presented at the Interspeech, Hyderabad, India, 2018.

[64] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[65] P.-S. Huang, X. He, J. Gao , L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, 2013, pp. 2333–2338.

[66] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proc. 23rd Int. Conf. World Wide Web*, Seoul, South Korea, 2014, pp. 373–374.

[67] H. Palangi *et al.*, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 4, pp. 694–707, Apr. 2016.

[68] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[69] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003.

[70] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," presented at the Int. Conf. Learn. Representations, Scottadale, AZ, USA, 2013.

[71] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," presented at the Int. Conf. Neural Inf. Process. Syst., Lake Tahoe, UT, USA, 2013.

[72] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Computat. Linguistics*, vol. 5, pp. 135–146, 2017.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," presented at the Conf. Comput. Vision Pattern Recognit., Miami, FL, USA, 2009.

[74] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, pp. 211–252, 2015.

[75] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *J. Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[76] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[77] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," presented at the Conf. Empirical Methods Natural Lang. Process., Doha, Qatar, 2014.

[78] A. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," in *Proc. Int. Conf. World Wide Web*, Florence, Italy, 2015, pp. 278–288.

[79] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Denver, 2015, pp. 912–921.

[80] W.-T. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proc. Assoc. Comput. Linguistics*, Baltimore, MD, USA, 2014, pp. 643–648.

[81] W.-T. Yih, M.-W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," in *Proc. Assoc. Comput. Linguistics*, Beijing, China, 2015, pp. 1321–1331.

[82] R. Kiros *et al.*, "Skip-thought vectors," presented at the Int. Conf. Neural Inf. Process. Syst., 2015.

[83] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," presented at the Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technologies, Atlanta, GA, USA, 2013.

[84] I. Goodfellow *et al.*, "Generative adversarial nets," presented at the Int. Conf. Neural Inf. Process. Syst., Monteral, Canada, 2014.

[85] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," presented at the Int. Conf. Learn. Representations, San Juan, Puerto Rico, 2016.

[86] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černocký, "iVector-based discriminative adaptation for automatic speech recognition," presented at the Autom. Speech Recognit. Understanding Workshop, Olomouc, Czech Republic, 2013.

[87] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. Workshop Autom. Speech Recognit. Understanding*, Olomouc, 2013, pp. 55–59.

[88] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Florence, Italy, 2014, pp. 225–229.

[89] S. Yella and A. Stolcke, "A comparison of neural network feature transforms for speaker diarization," presented at the Interspeech, Dresden, Germany, 2015.

[90] Q. Wang, C. Downey, L. Wan, P. Mansfield, and I. Lopez Moreno, "Speaker diarization with LSTM," presented at the Int. Conf. Acoust., Speech Signal Process., Calgary, Canada, 2018.

[91] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," presented at the Interspeech, Dresden, Germany, 2015.

[92] M. Wan, G. Degottex, and M. Gales, "Integrated speaker-adaptive speech synthesis," in *Proc. Autom. Speech Recognit. Understanding Workshop*, Okinawa, Japan, 2017, pp. 705–711.

[93] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[94] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," 2019, *arXiv:1806.04558*.

[95] R. Xia and Y. Liu, "DBN-ivector framework for acoustic emotion recognition," presented at the Interspeech, San Francisco, CA, USA, 2016.

[96] M. Sarma, P. Ghahremani, D. Povey, N. Goel, K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using DNNs," presented at the Interspeech, Hyderabad, India, 2018.

[97] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," 2020, *arXiv:2002.05039*.

[98] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[99] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," presented at the Interspeech, Lyon, France, 2013.

[100] N. Ström, "Speaker adaptation by modeling the speaker variation in a continuous speech recognition system," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, 1996, pp. 989–992.

[101] E. Variani, X. Lei, E. McDermott, I. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text- dependent speaker verification," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Florence, 2014, pp. 4052–4056.

[102] Z. Lin *et al.*, "A structured self-attentive sentence embedding," presented at the Int. Conf. Learn. Representations, Toulon, France, 2017.

[103] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," presented at the Interspeech, Hyderabad, India, 2018.

[104] G. Sun, C. Zhang, and P. Woodland, "Speaker diarisation using 2D self-attentive combination of embeddings," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., 2019, pp. 5801–5805.

[105] J. Portêlo, A. Abad, B. Raj, and I. Trancoso, "Secure binary embeddings of front-end factor analysis for privacy preserving speaker verification," presented at the Interspeech, Lyon, France, 2013.

[106] A. Nautsch *et al.*, "Preserving privacy in speaker and speech characterisation," *Comput. Speech Lang.*, vol. 58, pp. 441–480, 2019.

[107] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," presented at the Int. Conf. Mach. Learn., Bellevue, WA, USA, 2011.

[108] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," presented at the Int. Conf. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, 2012.

[109] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. Assoc. Comput. Linguistics*, Baltimore, MD, USA, 2014, pp. 721–732.

[110] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1473–1482.

[111] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional semantics in technicolor," in *Proc. Assoc. Comput. Linguistics*, Jeju Island, South Korea, 2012, pp. 136–145.

[112] S. Kottur, R. Vedantam, J. Moura, and D. Parikh, "Visual Word2Vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.

[113] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, HI, USA, 2017.

[114] P. Bachman, R. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," presented at the Conf. Neural Inf. Process. Syst., Vancouver, Canada, 2019.

[115] A. Lazaridou, N. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," presented at the Conf. North Amer. Chapter Assoc. Comput. Linguistics, Denver, CO, USA, 2015.

[116] A. Karpathy, A. Joulin, and F.-F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," presented at the Int. Conf. Neural Inf. Process. Syst., Montreal, Canada, 2014.

[117] H. Wu *et al.*, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 6602–6611.

[118] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," presented at the Eur. Conf. Comput. Vision, Munich, Germany, 2018.

[119] Y.-H. Tsai, P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," presented at the Int. Conf. Learn. Representations, Vancouver, Canada, 2018.

[120] T. Gupta, A. Schwing, and D. Hoiem, "ViCo: Word embeddings from visual co-occurrences," presented at the Int. Conf. Comput. Vision, Seoul, South Korea, 2019.

[121] D.-K. Nguyen and T. Okatani, "Multi-task learning of hierarchical vision-language representation," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[122] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," presented at the Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2013.

[123] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," presented at the Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2013.

[124] Y.-H. Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," presented at the Int. Conf. Comput. Vision, Venice, Italy, 2017.

[125] D. Golub, R. Martín-Martín, A. El-Kishky, and S. Savarese, "Leveraging pretrained image classifiers for language-based segmentation," in *Proc. Winter Conf. Appl. Comput. Vision*, Aspen, CO, USA, 2020, pp. 2010–2019.

[126] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proc. Int. Conf. Comput. Vision*, Santiago, 2015, pp. 4247–4255.

[127] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.

[128] A. Vaswani *et al.*, "Attention is all you need," presented at the Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017.

[129] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training," 2019, *arXiv:1908.06066*.

[130] W. Su *et al.*, "VL-BERT: Pre-training of generic visuallinguistic representations," 2019, *arXiv:1908.08530*.

[131] L. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[132] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," presented at the Int. Conf. Comput. Vision, Seoul, South Korea, 2019.

[133] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," presented at the Int. Conf. Mach. Learn. Comput., Shenzhen, China, 2019.

[134] H. Tan and B. Mohit, "LXMERT: Learning cross-modality encoder representations from transformers," presented at the Conf. Empirical Methods Natural Lang. Process., Hong Kong, 2019.

[135] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," presented at the Conf. Neural Inf. Process. Syst., Vancouver, Canada, 2019.

[136] S. Pramanik, P. Agrawal, and A. Hussain, "OmniNet: A unified architecture for multi-modal multi-task learning," 2019, *arXiv:1907.07804*.

[137] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," presented at the Assoc. Comput. Linguistics, Florence, Italy, 2019.

[138] B. Nojavanasghari, D. Gopinath, J. Koushik, B. T., and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. Int. Conf. Multimodal Int.*, 2016, pp. 284–288.

[139] H. Wang, A. Meghawat, L.-P. Morency, and E. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Tokyo, Japan, 2017.

[140] A. Anastasopoulos, S. Kumar, and H. Liao, "Neural language modeling with visual features," 2019, *arXiv:1903.02930*.

[141] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Proc. Eur. Conf. Comput. Vision*, Munich, Germany, 2018, pp. 575–589.

[142] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," 2015, *arXiv:1512.02167*.

[143] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[144] B. Zoph and Q. Le, "Neural architecture search with reinforcement learning," presented at the Int. Conf. Learn. Representations, Toulon, France, 2017.

[145] C. Liu *et al.*, "Progressive neural architecture search," in *Proc. Eur. Conf. Comput. Vision*, Munich, Germany, 2018, pp. 19–35.

[146] J.-M. Pérez-Rúa, M. Baccouche, and S. Pateux, "Efficient progressive neural architecture search," presented at the Brit. Mach. Vision Conf., Cardiff, U.K., 2019.

[147] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proc. ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, 2016, pp. 978–987.

[148] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," presented at the Int. Conf. Learn. Representations, San Diego, CA, USA, 2015.

[149] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," 2014, *arXiv:1410.5401*.

[150] Y. Zhu, O. Groth, M. Bernstein, and F.-F. Li, "Visual7W: Grounded question answering in images," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.

[151] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Lille, 2015, pp. 2048–2057.

[152] K. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Las Vegas, 2016, pp. 4613–4621.

[153] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Las Vegas, 2016, pp. 21–29.

[154] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vision*, Amsterdam, 2016, pp. 451–466.

[155] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. Int. Conf. Mach. Learn.*, New York, 2016, pp. 2397–2406.

[156] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, 2018, pp. 6077–6086.

[157] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang, "Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering," presented at the Assoc. Advancement Artif. Intell., New Orleans, 2018.

[158] W. Li *et al.*, "Object-driven text-to-image synthesis via adversarial training," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Long Beach, 2019, pp. 12174–12182.

[159] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," presented at the Int. Conf. Neural Inf. Process. Syst., Barcelona, 2016.

[160] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, 2017.

[161] H. Fan and J. Zhou, "Stacked latent attention for multimodal reasoning," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, 2018, pp. 1072–1080.

[162] A. Osman and W. Samek, "DRAU: Dual recurrent attention units for visual question answering," *Comput. Vision Image Understanding*, vol. 185, pp. 24–30, 2019.

[163] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," presented at the Int. Conf. Neural Inf. Process. Syst., 2017.

[164] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. González, "Gated multimodal units for information fusion," presented at the Int. Conf. Learn. Representations, Toulon, 2017.

[165] J.-H. Kim *et al.*, "Multimodal residual learning for visual QA," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Barcelon, 2016, pp. 361–369.

[166] H. Noh, P. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, 2016.

[167] J. Tenenbaum and W. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, pp. 1247–1283, 2000.

[168] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, 2017, pp. 1103–1114.

[169] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Las Vegas, 2016, pp. 317–326.

[170] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Proc. Int. Colloq. Automata, Lang. Program.*, Warwick, 2012, pp. 3–15.

[171] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, 2013, pp. 239–247.

[172] A. Fukui, D. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, 2016, pp. 457–468.

[173] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," presented at the Int. Conf. Learn. Representations, Toulon, 2017.

[174] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," presented at the Int. Conf. Comput. Vision, Venice, 2017.

[175] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.

[176] L. Tucker, "Some mathematical notes on three-mode factor analy," *Psychometrika*, vol. 31, pp. 279–311, 1966.

[177] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," presented at the Int. Conf. Comput. Vision, Venice, 2017.

[178] L. Lathauwer, "Decompositions of a higher-order tensor in block termspart II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 1033–1066, 2008.

[179] H. Ben-younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," presented at the Assoc. Advancement Artif. Intell., Honolulu, 2019.

[180] Z. Liu, Y. Shen, V. Lakshminarasimhan, P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," presented at the Assoc. Comput. Linguistics, Melbourne, 2018.

[181] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," presented at the Conf. Neural Inf. Process. Syst., Montreal, 2018.

[182] X. Wang *et al.*, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, 2019.

[183] X. He and L. Deng, "Deep learning for image-to-text generation: A technical overview," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 109–116, Nov. 2017.

[184] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.

[185] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," presented at the Conf. Comput. Vision Pattern Recognit., Boston, 2015.

[186] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," 2014, *arXiv:1412.6632*.

[187] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Boston, 2015, pp. 2422–2431.

[188] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," presented at the Int. Conf. Mach. Learn., Lille, 2015.

[189] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, 2017.

[190] Z. Gan *et al.*, "Semantic compositional networks for visual captioning," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, 2017.

[191] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, 2019.

[192] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," presented at the Eur. Conf. Comput. Vision, Amsterdam, 2016.

[193] K. Tran, X. He, L. Zhang, and J. Sun, "Rich image captioning in the wild," presented at the Conf. Comput. Vision Pattern Recognit. Workshop, Las Vegas, 2016.

[194] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Honolulu, 2017, pp. 955–964.

[195] D. Li, Q. Huang, X. He, L. Zhang, and M.-T. Sun, "Generating diverse and accurate visual captions by comparative adversarial learning," 2018, *arXiv:1804.00861*.

[196] A. Graves, "Generating sequences with recurrent neural networks," Generating sequences with recurrent neural networks, *arXiv:1308.0850*.

[197] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," presented at the Int. Conf. Mach. Learn., Lille, 2015.

[198] E. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov, "Generating images from captions with attention," presented at the Int. Conf. Learn. Representations, San Juan, 2016.

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

[200] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," presented at the Int. Conf. Neural Inf. Process. Syst., Montreal, 2015.

[201] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," presented at the Int. Conf. Mach. Learn., New York, 2016.

[202] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Barcelona, 2016, pp. 2234–2242.

[203] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6629–6640.

[204] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, 2017, pp. 2642–2651.

[205] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.

[206] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. Int. Conf. Comput. Vision*, Venice, Italy, 2017, pp. 5908–5916.

[207] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, pp. 1947–1962, 2019.

[208] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[209] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," presented at the Int. Conf. Learn. Representations, Vancouver, Canada, 2018.

[210] A. Dash, J. Gamboal, S. Ahmed, M. Liwicki, and M. Afzal, "TAC-GAN – Text conditioned auxiliary classifier generative adversarial network," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, HI, USA, 2017.

[211] M. Cha, Y. Gwon, and H. Kung, "Adversarial learning of semantic relevance in text to image synthesis," presented at the Assoc. Advancement Artif. Intell., Honolulu, HI, USA, 2019.

[212] X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[213] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[214] P. Welinder, "Caltech-UCSD birds 200," California Inst. Techn., Tech. Rep. CNS-TR-2010-001, 2010.

[215] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," presented at the Conf. Comput. Vision Pattern Recognit., New York, NY, USA, 2006.

[216] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp 740–755.

[217] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," presented at the Int. Conf. Neural Inf. Process. Syst., Barcelona, Spain, 2016.

[218] J. Johnson, A. Gupta, and F.-F. Li, "Image generation from scene graphs," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 1219–1228.

[219] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[220] T. Hinz, S. Heinrich, and S. Wermter, "Generating multiple objects at spatially distinct locations," presented at the Int. Conf. Learn. Representations, New Orleans, LA, USA, 2019.

[221] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.

[222] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Proc. Eur. Conf. Comput. Vision*, Amsterdam, The Netherlands, 2016, pp. 776–791.

[223] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[224] S. Nam, Y. Kim, and S. Kim, "Text-adaptive generative adversarial networks: Manipulating images with natural language," presented at the Conf. Neural Inf. Process. Syst., Montreal, Canada, 2018.

[225] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L. Jorio, and T. Fevens, "Dual adversarial inference for text-to-image synthesis," presented at the Int. Conf. Comput. Vision, Seoul, South Korea, 2019.

[226] F. Tan, S. Feng, and V. Ordonez, "Text2Scene: Generating compositional scenes from textual descriptions," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[227] S. Sharma, D. Suhubdy, V. Michalski, S. Kahou, and Y. Bengio, "ChatPainter: Improving text to image generation using dialogue," presented at the Int. Conf. Learn. Representations Workshop, Vancouver, Canada, 2018.

[228] A. El-Nouby *et al.*, "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction," presented at the Int. Conf. Comput. Vision, Seoul, South Korea, 2019.

[229] P. Cascante-Bonilla, X. Yin, V. Ordonez, and S. Feng, "Chat-crowd: A dialog-based platform for visual layout composition," presented at the Conf. North Amer. Chapter Assoc. Comput. Linguistics-Human Lang. Technologies, New, Orleans, LA, USA, 2018.

[230] Y. Chen, Z. Gan, Y. Li, J. Liu, and J. Gao, "Sequential attention GAN for interactive image editing via dialogue," presented at the Assoc. Adva. Artif. Intell., Honolulu, HI, USA, 2019.

[231] J.-H. Kim *et al.*, "CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication," presented at the Assoc. Comput. Linguistics, Florence, 2019.

[232] Y. Li *et al.*, "StoryGAN: A sequential conditional GAN for story visualization," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[233] Y. Li, M. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text," presented at the Assoc. Adv. Artif. Intell., New Orleans, LA, USA, 2018.

[234] Y. Balaji, M. Min, B. Bai, R. Chellappa, and H. Graf, "Conditional GAN with discriminative filter generation for text-to-video synthesis," presented at the Int. Joint Conf. Artif. Intell., Macao, China, 2019.

[235] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," presented at the Int. Conf. Neural Inf. Process. Syst., 2014.

[236] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," presented at the Int. Conf. Comput. Vision, Santiago, Chile, 2015.

[237] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, 2015, pp. 2953–2961.

[238] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? Reasoned visual dialog generation through adversarial learning," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.

[239] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, HI, USA, 2017, pp. 5415–5424.

[240] A. Das *et al.*, "Visual dialogue," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, HI, USA, 2017.

[241] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "GuessWhat?! Visual object discovery through multi-modal dialogue," presented at the Conf. Comput. Vision Pattern Recognit., 2017.

[242] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *Int. J. Comput. Vision*, vol. 127, pp. 398–414, 2019.

[243] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Explicit knowledge-based reasoning for visual question answering," presented at the Int. Joint Conf. Artif. Intell., Melbourne, Australia, 2017.

[244] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, "FVQA: Fact-based visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, pp. 2413–2427, Oct. 2018.

[245] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[246] D. Hudson and C. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.

[247] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Dont just assume; Look and answer: Overcoming priors for visual question answering," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, 2018.

[248] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," presented at the Conf. Neural Inf. Process. Syst., Montreal, Canada, 2018.

[249] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases in visual question answering," presented at the Conf. Neural Inf. Process. Syst. Vancouver, Canada, 2019.

[250] J.-Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," presented at the Int. Conf. Comput. Vision, Venice, Italy, 2017.
[251] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," presented at the Int. Conf. Learn. Representations, 2018.
[252] A. Singh *et al.*, "Towards vqa models that can read," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.
[253] D. Gurari *et al.*, "VizWiz grand challenge: Answering visual questions from blind people," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.
[254] X. Lin and D. Parikh, "Active learning for visual question answering: An empirical study," 2017, *arXiv:1711.01732*.
[255] I. Misra, R. Girshick, R. Fergus, M. Hebert, A. Gupta, and L. van der Maaten, "Learning by asking questions," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.
[256] K. Jedoui, R. Krishna, M. Bernstein, and L. Fei-Fei, "Deep Bayesian active learning for multiple correct outputs," 2019, *arXiv:1912.01119*.
[257] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," presented at the Assoc. Advancement Artif. Intell., New Orleans, LA, USA, 2018.
[258] R. Cadene, H. Ben-younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," presented at the Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA, 2019.
[259] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," presented at the Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA, 2016.
[260] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," presented at the Conf. North Amer. Chapt. Assoc. Comput. Linguistics, San Diego, CA, USA, 2016.
[261] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proc. Int. Conf. Comput. Vision*, Venice, Italy, 2017, pp. 804–813.
[262] J. Johnson, B. Hariharan, L. van der Maaten, F.-F. Li, C. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," presented at the Conf. Comput. Vision Pattern Recognit., Honolulu, HI, USA, 2017.
[263] J. Johnson *et al.*, "Inferring and executing programs for visual reasoning," presented at the Int. Conf. Comput. Vision, Venice, Italy, 2017.
[264] R. Hu, J. Andreas, T. Darrell, and K. Saenko, "Explainable neural computation via stack neural module networks," presented at the Eur. Conf. Comput. Vision, Munich, Germany, 2018.
[265] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," presented at the Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, 2018.
[266] D. Hudson and C. Manning, "Compositional attention networks for machine reasoning," presented at the Int. Conf. Learn. Representations, Vancouver, Canada, 2018.
[267] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," presented at the Conf. Neural Inf. Process. Syst., Montreal, Canada, 2018.
[268] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural-symbolic models for interpretable visual question answering," presented at the Int. Conf. Mach. Learn., Stockholm, Germany, 2018.
[269] J. Mao, C. Gan, P. Kohli, J. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," presented at the Int. Conf. Learn. Representations, New Orleans, LA, USA, 2019.
[270] A. Santoro *et al.*, "A simple neural network module for relational reasoning," presented at the Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017.
[271] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," presented at the Eur. Conf. Comput. Vision, Amsterdam, The Netherlands, 2016.
[272] Google, "Freebase data dumps," 2015. [Online]. Available: https://developers.google.com/freebase
[273] J. Han, Z. Zhang, R. Zhao, and B. Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," 2019, *arXiv:1912.01119*.
[274] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, 2019.

Chao Zhang received his B.E. and M.S. degrees in 2009 and 2012, respectively, from the Department of Computer Science & Technology, Tsinghua University, and his Ph.D. in 2017 from the Cambridge University Engineering Department. He is currently an Advisor for the JD.com speech team and a Research Associate of speech and language processing at the University of Cambridge.

Zichao Yang received his Ph.D. in Computer Science from Carnegie Mellon University. He is currently a Quantitative Researcher at Citadel. His research interests include machine learning, deep learning, and their applications to natural language processing and computer vision. He has published dozens of papers in NeurIPS, ICML, CVPR, ICCV, EMNLP, NAACL, etc.

Xiaodong He (Fellow, IEEE) holds a B.S. degree from Tsinghua University, M.S. degree from the Chinese Academy of Sciences, and Ph.D. degree from the University of Missouri, Columbia. He is currently the Deputy Managing Director of JD AI Research and the Head of the Deep Learning, NLP, and Speech Lab. He is also an Affiliate Professor of ECE at the University of Washington, Seattle. His research interests mainly focus on deep learning, natural language processing, speech recognition, computer vision, information retrieval, and multimodal intelligence. He has held editorial positions for IEEE TASLP, JSTSP, SPM, SPL and for the *Transactions of the ACL* (TACL). He has also served in the organizing committees/program committees of major speech and language processing conferences. He was a member of the IEEE SLTC for the term of 2015 to 2017 and the Chair of the IEEE Seattle Section in 2016–2017.

Li Deng (Fellow, IEEE) has been the Chief Artificial Intelligence Officer of Citadel since May 2017. Prior to joining Citadel, he was the Chief Scientist of AI, the founder of the Deep Learning Technology Center, and a Partner Research Manager at Microsoft and the Microsoft Research of Redmond, from 2000 to 2017. Prior to joining Microsoft, he was an Assistant Professor, from 1989 to 1992, Tenured Associate from 1992 to 1996 and Full Professor, from 1996 to 1999 at the University of Waterloo, Ontario, Canada. He also held faculty or research positions at the Massachusetts Institute of Technology, Cambridge from 1992 to 1993. Advanced Telecommunications Research Institute, Kyoto, Japan, from 1997 to 1998, and HK University of Science and Technology, Hong Kong, in 1995. He is a Fellow of the Academy of Engineering of Canada, Washington State Academy of Sciences, Acoustical Society of America, and International Speech Communication Association. He has also been an Affiliate Professor at the University of Washington, Seattle.

Prof. Deng was an elected member of the Board of Governors of the IEEE Signal Processing Society and was the Editor-in-Chief of IEEE SIGNAL PROCESSING MAGAZINE AND IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2008 in 2014), for which he received the IEEE SPS Meritorious Service Award. In recognition for this pioneering work in the speech recognition industry using large-scale deep learning, he received the 2015 IEEE SPS Technical Achievement Award for "Outstanding Contributions to Automatic Speech Recognition and to Deep Learning." He also received dozens of best paper and patent awards for his contributions to artificial intelligence, machine learning, information retrieval, multimedia signal processing, speech processing and recognition, and human language technology. He is an author and a co-author of six technical books on deep learning, speech processing, pattern recognition, machine learning, and natural language processing (Springer, June, 2018).