Multimodal Intelligence: Representation Learning, Information Fusion, and Applications

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

***Abstract*—Deep learning has revolutionized speech recognition, image recognition, and natural language processing since 2010, each involving a single modality in the input signal. However, many applications in artificial intelligence involve more than one modality. It is therefore of broad interest to study the more difficult and complex problem of modeling and learning across multiple modalities. In this paper, a technical review of the models and learning methods for multimodal intelligence is provided. The main focus is the combination of vision and natural language, which has become an important area in both computer vision and natural language processing research communities.**

This review provides a comprehensive analysis of recent work on multimodal deep learning from three new angles — learning multimodal representations, the fusion of multimodal signals at various levels, and multimodal applications. On multimodal representation learning, we review the key concept of embedding, which unifies the multimodal signals into the same vector space and thus enables cross-modality signal processing. We also review the properties of the many types of embedding constructed and learned for general downstream tasks. On multimodal fusion, this review focuses on special architectures for the integration of the representation of unimodal signals for a particular task. On applications, selected areas of a broad interest in current literature are covered, including caption generation, text-to-image generation, and visual question answering. We believe this review can facilitate future studies in the emerging field of multimodal intelligence for the community.

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

I. Introduction

SIGNIFICANT progress has been made in the field of machine learning in the past years due to the rapid development of deep learning [1]–[6]. Dating back to the dramatic increase in the accuracy of large-scale automatic speech recognition (ASR) using fully connected deep neural networks (DNN) and deep auto-encoders around 2010 [7]–[17], and followed by a set of breakthroughs 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 [23]–[25] around 2014, a set of major milestones have been achieved in pattern recognition with single input modality. Subsequently, in natural language processing (NLP), recurrent neural network (RNN) based semantic slot filling methods [26] achieved new state-of-the-art in spoken language understanding, and RNN-encoder-decoder models with attention mechanism [27], also referred to as sequence to sequence models [28], produced superior performance in machine translation in an end-to-end fashion [29], [30]. For other NLP tasks without much training data, such as question answering (QA) and machine reading comprehension, generative pre-training that transfers parameters from a language model (LM) pre-trained on a large out-of-domain data set using unsupervised or self learning then followed by fine-tuning on small in-domain data sets, achieved record-breaking results over a set of tasks [31]–[33].

Despite the advances in vision, speech, and language processing, many problems in artificial intelligence involve more than one modality, such as an intelligent personal assistant (IPA) that should understand human communicative intentions embedded not only in spoken language, but also in body and pictorial languages [34]. Therefore, it is of broad interests to study the modeling and learning approaches across multiple modalities [35]. Benefiting from the advances in image processing and language understanding [36], a set of tasks that combine both image and text have drawn much attention, which include visual grounding tasks like referring expression understanding and phrase localization [37]–[39], image captioning [40]–[42], visual QA (VQA) [43]–[45], text-to-image generation [46]–[48], and visual-language navigation [49] *etc*. In these tasks, natural language plays a key role in helping the machine to understand the content of the images, where understand means to capture the underlying correlations between the semantics embedded in language with the visual features obtained from the images. In addition to text, vision can be combined with speech as well. Such tasks include audio-visual speech recognition [50]–[52], speaker recognition [53]–[55], as well as speech diarisation [56], [57], separation [58], [59] and enhancement [60], which mostly focused on the use of visual features to improve the robustness of the audio-only methods.

In this paper, a technical review of the models and learning methods for multimodal intelligence is provided. The main focus is the combination of CV and NLP, which has become an important area for both research communities covering many different tasks and technologies. To provide a more structured perspective, we organize the methods selected in this technical review according to three key topics: representation, fusion, and applications.

- Learning representations for the input data is a core problem for deep learning. For multimodal tasks, collecting paralleled data across all modalities can be quite difficult and leveraging pre-trained representations with desired properties, such as suitable for zero-shot or few-shot learning, is often an effective solution to the issue. Both

C. Zhang and X. He were with JD AI Research. C. Zhang was also with the Department of Engineering, University of Cambridge, UK.

Z. Yang and L. Deng were with Citadel LLC.

supervised and unsupervised training based multimodal representation learning methods are reviewed.
- The fusion of the features or representations of the single modalities is undoubtedly a centric problem of any multimodal task. Different from previous studies that often categorise the related work into early, middle and late stage methods based on the stage that fusion happens in the procedure, we classify them according to the actual operation used in the fusion, such as attention and bilinear pooling, since it becomes difficult to classify some recent complex approaches into stages.
- Three types of applications are reviewed in this paper, namely image captioning, text-to-image synthesis and VQA. This is to give an idea how representation learning and fusion can be applied to specific tasks, and to provide a viewpoint of the situation of the current development of the multimodal applications, especially those integrating vision with natural languages. Visual reasoning methods for VQA are also discussed in the end.

This paper is organised as follows. Section II reviews the recent progress on developing representations for single or multiple modalities. Section III introduces the commonly used fusion methods, particularly attention and bilinear pooling. Applications including caption generation, text-to-image synthesis, VQA, and visual reasoning are introduced in Section IV, followed by conclusions.

II. Representations

Deep learning, as a special area in representational learning, studies the use of artificial neural networks (ANNs) with many hidden layers to automatically discover the representations or features suitable for specific tasks from the raw data [61]. In practice, it is often found that better representations can simplify the subsequent learning tasks and therefore has a great value. Over the past decade, it becomes feasible to learn effective and robust representations for single modalities, such as text [31]–[33], [62]–[69] and image [19]–[25], due to the availability of large data and development of deep learning. For multimodal representations, though attracting more and more attentions, it still remains a challenging problem due to the complex cross-modal interactions and possible mismatch between the training and test data of each modal.

In this section, commonly used types of single modal representations, such as text and image, are first reviewed which often serve as cornerstones for learning multimodal representations. Afterwards, both supervised and unsupervised methods for learning a joint representation space for multiple modalities are introduced. To empower the model to handle data samples with some missing modality, zero-shot learning problem is studied to increase the similarity of the representational spaces across the involved modalities. At last, inspired by the great success of adapting pre-trained LMs for downstream tasks in NLP, methods that leverage large unimodal data sets to improve the learning of multimodal representations are also discussed.

A. Unimodal Embeddings

The representations obtained using ANN models are often distributed, which entails that elements composing the representations can be set separately to allow more concepts to be encoded efficiently in a relatively low-dimensional space [65]. This can be compared with the symbolic representations, such as the one-hot encoding that uses an element with value one to indicate the presence of the associated symbol or category, and value zero for the rest elements. In deep learning, the term embedding often refers to a mapping from a one-hot encoding representing a word or an image category to a distributed representation as a vector of real-valued numbers.

1) Visual representation: The image embeddings can be acquired as the output values from the final CNN layers from models that classify images into categories, such as AlexNet [19], VGG nets [20], and residual neural network (ResNet) [21]. AlexNet is a model with five CNN layers with rectified linear unit (ReLU) activation functions whose kernel sizes are 11×11, 5×5, and 3×3. A VGG net often has 16 or 34 CNN layers, with all of them using very small 3×3 kernels. ResNet can have a depth of up to 152 layers, mostly with 3×3 kernels, due to the invention of residual connections. Comparing to the aforementioned models, GoogLeNet has a more different structure formed by stacking multiple Inception structures [22]. The naïve Inception structure is a concatenation of CNN layers with 1×1, 3×3, and 5×5 sized kernels, and a max pooling layer with 3×3 kernels, and can be viewed as a sparsely connected convolutional architecture to reduce overfitting and computational cost. Later versions of the Inception models improve the structures further by factorizing the kernels and adding residual connections. AlexNet, GoogLeNet, and ResNet are the winners of the 2012, 2014, and 2015 ImageNet Large Scale Visual Recognition Competition for image classification respectively [70], [71]. Alternatively, features with more direct relationships with the semantics can be used as visual embeddings, such as convolutional features and the associated class labels from selected regions found by object detection models, such as region with CNN features (R-CNN) [23], Fast R-CNN [24], and Faster R-CNN [25] *etc*.

2) Language representations: Text embeddings can be derived from a neural network language model (NNLM) [66], which estimates the probability of a text sequence by factorizing it into word probabilities based on the chain rule of probability. A feedforward neural network with a linear projection layer and a non-linear hidden layer is often used as an implementation of an n-gram LM, which takes the previous $n-1$ words as the input to predict the probability of the current word. Each word is presented in one-hot encoding based on the vocabulary and converted into a real-valued vector, the word embedding, using the projection layer. An improved NNLM is to replace the feedforward model with an RNN, such as a long short-term memory (LSTM) [72] or gated recurrent unit (GRU) [73] model, which allows the use of information from all past words stored in a fixed-length recurrent vector when predicting the current word. Apart from NNLMs, continuous bag-of-words model (CBOW) and skip-gram model are two

simple feedforward structures that learn word embedding either by predicting the current word based on the past and future context words or *vice versa* [67]. The method of global vectors (GloVe) shows that effective text-embedding can be learnt using a global log-bilinear regression model based on the co-occurrence counts of words [74]. Meanwhile, a serious of deep structured semantic models (DSSM) were proposed since 2013 for sentence level embedding learning through optimizing semantic-similarity driven objectives, with various neural network structures in a pseudo-simense network setting [62]–[64], [75]–[78].

More recently, in order to transfer to use in downstream natural language understanding tasks without much training data, studies focus on learning general text embeddings by predicting word probabilities using NNLMs with complex structures on a large text corpus. Embeddings from language models (ELMo) uses a combined embedding from multiple layers of bidirectional LSTMs for forward and backward directions [31]. Generative pre-training (GPT) and bidirectional encoder representations for Transformers (BERT) use the decoder and encoder part of the Transformer model to estimate the probability of the current subword unit [32], [33]. Other technologies, such as masked language model and multi-task training are used in these methods nowadays [33]. Besides word and subword levels, text embedding can be learnt at phrase, sentence, and even paragraph levels, such as the skip-though vectors that extends the skip-gram method to the sequence-to-sequence framework [28], [79]. It uses two decoders to predict the previous and next sentences given the embedding of the current sentence generated by the encoder.

3) Vector arithmetic for word and image embeddings: It is well-known that word embeddings can learn not only syntactic but also the semantic regularities. A famous example showed vector("King")−vector("Man")+vector("Woman") results in a vector closest to vector("Queen") where vector(·) denotes the vector representation of a word learnt by a RNN LM [80]. Similar phenomenon has been observed for vision embeddings. It was shown that when using a generative adversarial network (GAN) [81], there exists a similar vector arithmetic that the representation of an image with a man wearing glasses subtracted by that of a man without glasses and finally add the representation of a woman without glasses will lead to the representation of a woman wearing glasses [82]. This reveals GAN can capture image representation that distangles the concept of gender from the concept of wearing glasses. Such encouraging progress in text and image representations encouraged further studies on the joint representations of these two modalities. More details about GAN based image generation can be found later in Section IV-B.

B. Multimodal Representations

Although significant progress has been made in the learning of representations for vision or language, it is theoretically insufficient to model a complete set of human concepts using only unimodal data. For example, the concept of "beautiful music" is clearly grounded in the auditory perceptron and one can be struggled to describe this by natural language or other approaches. Therefore, it is important to learn a joint embedding to leverage the complementarity from multimodal data to represent the concepts better. Both supervised and unsupervised training approaches are of broad interest and can be applied to tasks with different data availability. Meanwhile, by assuming the corresponding representations to have similar neighbourhood structures across modalities, the representation of a concept with zero training sample in one modal can be found based on its representations grounded in other modalities which have training data. For instance, when using zero-shot training for image labelling, the closest word vectors can be retrieved as labels by projecting images of objects unseen in the training set onto the linguistic space. More recently, inspired by Transformer and BERT from NLP, it becomes increasingly popular to apply these models to develop better bimodal representations combining vision and language.

1) Unsupervised training methods: Joint embeddings for multimodal data can be learnt by simply reconstructing the raw input using multiple streams of deep Boltzmann machines or autoencoders with a shared layer as the shared representation space [83]–[85]. Alternatively, with the development of methods for single modal representations, the shared representation space can be constructed based on those of the involved single modalities. For example, in [86], Fang *et al*. propose a deep multimodal similarity model (DMSM), which extended the text-modal DSSM to learning embedding representations of text and image in an unified vector space. [85], [87] perform simple fusion of the word and image embeddings with addition or concatenation. [88] learns to increase the similarity between corresponding Skip-Gram word embedding and AlexNet derived image features. [89], [90] maximize the correlation and mutual information between embeddings of different modalities respectively. [91] modifies the distance between CBOW word embeddings according to the similarities between their visual instantiations, which are found by clustering abstract scenes in an unsupervised way.

Further studies found correlating image regions/fragments with sentence fragments or attribute words generates fine-grained multimodal embeddings [92], by finding the alignments of the image and sentence fragments automatically. [93] unifies the embeddings of concepts at different levels, including objects, attributes, relations and full scenes. [94] proposed a stacked cross attention network (SCAN) to learning fine-grained word and image-object aligned embedding for image-text matching. [48] employs a deep attentional multimodal similarity model (DAMSM) extending DMSM with attention models to measure the similarity between image sub-regions and words as an additional loss for text-to-image generation.

2) Supervised training methods: Supervisions can be used to improve multimodal representation learning. [95] factorizes the representations into two sets of independent factors: multimodal discriminative factors for supervised training and intra-modality generative factors for unsupervised training. The discriminative factors are shared across all modalities and are useful for discriminative tasks, whereas the generative factors can be used to reconstruct missing modalities. With detailed text annotations, [96] proposed to learn word embeddings from their visual co-occurrences (ViCo) when applying to the same

natural scene image or image region. ViCo is found to be complementary to the GloVe embedding by better representing similarities and differences between visual concepts that are difficult to obtain from text corpora alone. [97] applies multiple supervised training tasks to different layers of the vision-language encoder. The order of the training tasks is arranged following the idea of curriculum learning to increase the complexity of training objective step-by-step.

3) Methods for zero-shot learning: Zero-shot learning often applies to vision related tasks due to the difficulty to acquire sufficient labelled images for training for all possible object categories. Not all types of multimodal representations are suitable for zero-shot learning since they may require pair-wised data from both modalities to present at the same time. Here we review methods that rely on extra language source to remedy this issue.

Deep learning based zero-shot learning started by training a linear mapping layer between different pre-trained embeddings [98], [99]. The deep visual-semantic embedding (DeViSE) model is built upon Skip-Gram word embedding and AlexNet visual features and allows both pre-trained models to be jointly trained with the linear mapping layer [99]. It achieved a large-scale test with 1000 seen classes and 2000 unseen classes. Better representations could be learnt when correlated autoencoders are used to reconstruct the representations for each modality, which improves one-shot and few-shot image retrieval performance comparing to DeViSE [100]. Richer information source can be used for both modalities, including words selected from Wikipedia articles and features derived from multiple CNN layers [101]. Rather than direct text attribute input, sentence embedding generated by recurrent models can be used as the text interface for zero-shot learning to achieve competitive results [102]. Moving beyond empirical findings, recent study analyzed the properties of deep learning based cross-modal mapping using a similarity measure [103].

4) Transformer based methods: Transformer is a prevalent sequence-based encoder-decoder model formed by stacking many blocks of feedforward layers with multi-head self-attention models [104]. The parameters in all blocks shared across time similar to the time-delayed neural networks [105] and quasi-RNN [106] without an explicit temporal order. Compared with RNN based encoder-decoder models [27], it can have higher training efficiency due to the additional parallel degree across-time and superior performance on longer sequences benefited from the removing of first-order Markovian assumption imposed to the RNNs. BERT, the encoder part of Transformer pre-trained on a large text corpus as a masked LM, becomes a standard choice for word piece embeddings for downstream tasks, particularly since it utilizes both past and future information easily. It is natural to generalize the text-only BERT to cover images as well that can be used as the pre-trained multimodal embeddings.

A straightforward way to extend the unimodal BERT to bimodal, is to include new tokens to indicate the input of visual features, such as Unicoder-VL [107], VL-BERT [108], VisualBERT [109], VideoBERT [110], and B2T2 [111]. LXMERT [112], ViLBERT [113], and OmniNet [114] modify the Transformer model by introducing an extra encoder or attention structures for visual features. More details about the modified structures can be found from Section III-B. Furthermore, recent NLP study found that multitask training can improve the generalization ability of the BERT representations [115]. Most of the aforementioned bimodal BERT style models adopt multitask training to improve their performance on downstream tasks like VQA, image and video captioning *etc*. Although it would be useful to rigorously compare the performance of these models to understand the impact of different design choices, it is hard to do so since different amount of parameters and pre-training data are used across papers.

III. Fusion

Fusion is a key research problem in multimodal studies, which integrates information extracted from different unimodal data into one compact multimodal representation. There is a clear connection between fusion and multimodal representation. We classify an approach into the fusion category if its focus is the architectures for integrating unimodal representations for particular a task.

Traditionally, fusion methods are divided based on the stage it appears in the procedure. Early fusion, or feature-level fusion, directly combines the features extracted from each type of unimodal data to stress the intra-modality interactions and can cause the inter-modality interactions to be suppressed. Late fusion, on the other hand, refers to model-level fusion that builds a separate model for each modality and combines their output [116]–[120]. The late fusion methods are strong in modelling intra-modality interactions with the modality-specific models but may suffer from the limited power of simple output value combination since the inter-modality interactions are rather complex. Recent studies focus on the intermediate or middle-level methods that allows fusion to happen at multiple layers of a deep model.

In this section, a review on intermediate fusion is focused – not only as it is more flexible, but also because the boundaries between stages are less clear due to the use of unimodal features derived from pre-trained backbone models. Three types of methods mostly used to fuse text with image features are included: simple operation-based, attention-based, as well as tensor-based methods.

A. Simple Operation-based Fusion

In deep learning, vectorized features from different information sources can be integrated using a simple operation, such as concatenation or weighted sum, which often has only a few or even no parameter associated since the joint training of deep models can adapt the layers for high-level feature extractions to adjust for the required operation.

- Concatenation can be used to combine either low-level input features [120]–[122] or high-level features extracted by the pre-trained models [122]–[124].
- For weighted sum with scalar weights, an iterative method is proposed [125] that requires the pre-trained vector representations for each modality to have the same number of elements arranged in an order that is suitable

for element-wise addition. This is often achieved by jointly training a fully connected layer for dimension control and reordering for each modality, together with the scalar weights for fusion.

A recent study [126] employs neural architecture search with progressive exploration [127]–[129] to find suitable settings for a number of fusion functions. Each fusion function is configured by which layers to fuse and whether to use concatenation or weighted sum as the fusion operation. Other weak functions can also be used to fuse multiple layers from each modality [130].

B. Attention-based Fusion

Attention mechanism is widely used for fusion, which often refers to weighted sum a set of vectors using scalar weights dynamically generated by a small attention model at each time-step [130], [131]. Multiple glimpses (output heads) are often used by the attention model to generate multiple sets of dynamic weights for the summation, whose resulted values can be concatenated to reserve more information. When applying attention mechanism to an image, image feature vectors relevant to different regions are weighted differently to produce an attended image vector.

1) Image attention: [132] extends an LSTM model for text question processing with an image attention model conditioned on the previous LSTM hidden state, whose input is a concatenation of the current word embedding with the attended image feature. The final LSTM hidden state is regarded as the fused multimodal representation to predict the answer for pointing and grounded VQA. The attention model for sequence-based encoder-decoder model is used to attend to the image features for image captioning [133]. Further for VQA, attention model conditioned on both image and query feature vectors is applied to pinpoint the image regions relevant to the answer [134]. Similarly, stacked attention networks (SANs) are proposed to use multiple layers of attention models to query an image multiple times to infer the answer progressively by simulating a multi-step reasoning procedure [135]. At each layer, a refined query vector is generated and send to the next layer by adding the previous query vector to the attended image vector produced using the current attention model. Spatial memory network (SMem) is a multi-hop method for VQA, which aligns words to image regions in the first hop and performs image attention based on the entire question in the second hop to derive the answer [136].

In [137], dynamic memory network (DMN) is augmented to use separate input modules to encode the question and image, which uses attention based GRUs to update episodic memory iteratively to retrieve the required information. Bottom-up and top-down attention method (Up-Down), as its name suggested, simulates human visual system using a combination of two visual attention mechanisms [138]. The bottom-up attention mechanism proposes a set of salient image regions found by a Faster R-CNN, and the top-down attention mechanism uses a concatenation of visual and linguistic features to estimate the attention weights and produce the attended image feature vector for image captioning or VQA. The attended image feature vector can be fused with the linguistic feature again using an element-wise product. Complementary image features derived from different models, such as ResNet and Faster R-CNN, are used for multiple image attention mechanisms [139]. Moreover, the reverse of image attention that generates attended text feature with image and text input is used for text-to-image synthesis in [48] and [140].

2) Image and text co-attention: Different from the aforementioned image attention methods, co-attention mechanism has a symmetric attention structure to generate not only an attended image feature vector, but also an attended language vector [141]. The parallel co-attention uses a joint representation to derive the image and language attention distributions simultaneously; alternating co-attention, on the other hand, has a cascade structure that first generates the attended image vector using the linguistic features, followed by the attended language vector generated using the attended image vector.

Similar to the parallel co-attention, dual attention network (DAN) estimates attention distributions for image and language simultaneously to derive their attended feature vectors [142]. The attention models are conditioned on both feature and memory vectors of the relevant modality. This is a key difference to co-attention since the memory vectors can be iteratively updated at each reasoning step by repeating the DAN structure. The memory vectors can be either shared for VQA or modality-specific for image-text matching. Stacked latent attention (SLA) improves SAN by concatenating the original attended image vector with values from earlier layers of the attention model to retain the latent information from intermediate reasoning stages [143]. A parallel co-attention like twin stream structure is also included to attend to both image and language features that also allows to reason iteratively using multiple SLA layers. Dual recurrent attention units (DRAU) implements the parallel co-attention structure with LSTM models for text and image to attend to each input location of the representations obtain by convolving image features with a stack of CNN layers [144]. To model high-order interactions between modalities, high-order correlations between two data modalities can be computed as the inner product of two feature vectors and used to construct high-order attention models to derive the attended feature vectors for both modalities [145].

3) Attention in bimodal Transformer: Recall Section II-B4, the bimodal extensions to BERT rely on different tokens to indicate whether a vector is a word piece or an image, and the attention models fuse images with words in bimodal input sequences [107]–[111]. OmniNet uses the gated multi-head attention model in each decoder block to fuse the vectors from the other modalities with that produced for the current modality by the previous layers in the the block [114]. LXMERT uses independent encoders to learn the intra-modality features for each modality, and a cross-modality encoder sitting above them to learn the cross-modality features using extra cross-attention layers [112]. ViLBERT extends BERT to include two encoder streams to process visual and textual inputs separately, which can interact through parallel co-attention layers [112].

4) Other attention like mechanisms: Gated multimodal unit is a method that can be viewed as the attention of image and

text based on gating [146]. It performs weighted sum of visual and textual feature vectors based on dimension-specific scalar weights generated dynamically by the gating mechanism. Similarly, element-wise multiplication can be used to fuse visual and textual representations, which is used to create the building blocks of a multimodal residual network (MRN) based on deep residual learning [147]. Dynamic parameter prediction network (DPPnet) uses a dynamic weight matrix to transform the visual feature vectors, whose parameters are dynamically generated by hashing the text feature vector [148].

C. Bilinear Pooling-based Fusion

Bilinear pooling is a method often used to fuse a visual feature vector with a textual feature vector into a joint representation space by computing their outer product, which allows a multiplicative interaction between all elements in both vectors and is also termed as second order pooling [149]. Comparing to simple vector combination operations (assuming each vector has n elements), such as weighted sum, element-wise multiplication, or concatenation that result in n or $2n$ dimensional representations, bilinear pooling leads to an n^2 dimensional representation by linearizing the outer product resulted matrix into a vector and is therefore more expressive. The bilinear representation is often linearly transformed into an output vector using a two-dimensional weight matrix, which is equivalent to use a three-dimensional tensor operator to fuse the two input feature vectors. Each feature vector can be extended with an extra value one to reserve input single modal features in the bilinear representation via outer product [150]. However, given its high dimensionality, typically on the order of hundreds of thousands to a few million, bilinear pooling often requires decomposing the weight tensor to have the associated model to be trained properly and efficiently.

1) Factorization for bilinear pooling: Since bilinear representations are found to be closely related to the polynomial kernels, different low-dimensional approximations can be used to acquire compact bilinear representations [151]. Count Sketches and convolutions can be used to approximate the polynomial kernels [152], [153] that leads to the multimodal compact bilinear pooling (MCB) method [154]. Alternatively, by enforcing a low rank to the weight tensor, multimodal low-rank bilinear pooling (MLB) factorizes the three-dimensional weight tensor for bilinear pooling into three two-dimensional weight matrices [155]. More precisely, the visual and textual feature vectors are linearly projected to low-dimensional modality-specific factors by the two input factor matrices, which are then fused using element-wise multiplication followed by a linear projection with the third matrix, the output factor matrix. Multimodal factorized bilinear pooling (MFB) modifies MLB by using an extra operation to pool the element-wise multiplication results by summing the values within each non-overlapped one-dimensional window [156]. Multiple MFB models can be cascaded to model high-order interactions between input features and is called multimodal factorized high-order pooling (MFH) [157].

MUTAN, a multimodal tensor-based Tucker decomposition method, uses Tucker decomposition [158] to factorize the original three-dimensional weight tensor operator with a small-dimensional core tensor and the three two-dimensional weight matrices used by MLB [159]. The core tensor models the interactions across modalities. Comparing to MUTAN, MCB can be seen as MUTAN with fixed diagonal input factor matrices and a sparse fixed core tensor, while MLB is MUTAN with the core tensor set to identity. Recently, BLOCK, a block superdiagonal fusion framework is proposed to use block-term decomposition [160] to compute bilinear pooling [161]. BLOCK generalizes MUTAN as a summation of multiple MUTAN models to provide a richer modeling of interactions between modalities. The MUTAN core tensors can be arranged as a superdiagonal tensor, similar to the submatrices of a block diagonal matrix. Furthermore, bilinear pooling can be generalized to more than two modalities, such as [150] and [162] that use outer products to model the interactions among the representations for video, audio, and language.

2) Bilinear pooling and attention mechanism: Bilinear pooling can be used along with attention mechanism. MCB/MLB fused bimodal representation can be used as the input feature of an attention model to derive the attended image feature vector, which is fused with the textual feature vector using MCB/MLB again to form the final joint representation [154], [155]. MFB/MFH can be used for alternating co-attention to learn the join representation [156], [157]. Bilinear attention network (BAN) uses MLB to fuse image and text to produce a bilinear attention map as the attention distributions, which is used as the weight tensor for bilinear pooling to fuse the image and text features again [163].

IV. Applications

In this section, selected applications for multimodal intelligence that combine vision and language are discussed, which include image captioning, text-to-image generation, and VQA. It is worth noting that there are other applications, such as text-based image retrieval [94], [164], [165], and visual-and-language navigation (VLN) [166]–[174], that we have not included in this paper due to space limitation.

- Caption generation is a task that aims to automatically generate a natural language description of an image. It requires a level of image understanding beyond normal image recognition and object detection.
- A reverse of caption generation is text-to-image synthesis, which often generates image pixels according to a description sentence or some key words provided by human.
- VQA is related to caption generation, which often takes an image as the input and a free-form, open-ended natural language question about the image, to output a classification result as the output of the answer. Natural language understanding is required as the questions are in free form. Other capabilities such as knowledge based reasoning and common-sense reasoning can be important since the questions are open-ended.
- Visual reasoning can be included in all of the aforementioned tasks. Visual reasoning methods for VQA are reviewed in the end.

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

A. Caption Generation

Image captioning [175] requires to generate a description of an image and is one of the earliest task that studies multi-modal combination of image and text. We mainly review the deep learning based methods for caption generation. Image captioning, such as [40], [86], [176], divide the task into several sub-tasks and generate caption in a step-by-step manner. Authors in [86] first trained a deep CNN model to detect the words from images, then built a log-linear language model to compose the words into sentences. Similarly, [176] fed the image feature into a log-linear language model to generate sentences. In contrast, [40] tried to find the exact matching of objects in images and words in sentences to determine if an image and a sentence match with each other.

Similar to the RNN-based encoder-decoder methods for machine translation [27], [177]–[179] propose to generate captions from images in an end-to-end manner via the encoder-decoder architecture. In those models, a CNN, typically pre-trained on ImageNet [70], encoded the image into a continuous vector, which is then fed into a RNN/LSTM decoder to generate the caption directly. Those works all followed the same architecture, but varied slightly the choice of CNN architecture and how the image vector was fed into the decoder. Though powerful and convenient, the encoder-decoder architecture lacks to ability capture the fine grained relationship between objects in images and words in sentences. To overcome this, attention-based encoder-decoder model [180] was proposed and has become the standard benchmark for this task since then. In the attention encoder-decoder model, before generating the next word, the decoder first calculates the matching score (attention) with objects in the image, then conditions on the weighted image feature to generate the next token. There has been lots of work that tried to improve the attention model by incorporating more structures. For example, [181] added a gate at every decoding step to determine if the next word should be generated using image information; [182] combined detected words and image features as inputs to the decoder network. More recently, there has been a lot works that add more structure/knowledge from either image [138] or text side [183]. Specifically, [138] used an object detector to localize the features for image object and then generates the caption based on the localized features. It improved the previous state of art model by a large margin in a variety of evaluation metrics.

Image captions with richer information could be generated when incorporated with external knowledge. For example, based on an model that can recognize celebrities [184], a CaptionBot app is developed which can not only describe the facts (such as activities) in a picture, but also describe who is doing that if the person in the picture is recognized [185]. Further, beside simply generating a factual description of the image, other approaches were also proposed for explicitly controlling the style [186], semantic content [182], and diversity [187] of the generated caption.

B. Text-to-Image Synthesis

Text-to-image synthesis or generation that relies on natural language to control image generation, is a fundamental problem in computer vision. It is considered as a difficult problem since it least involves two tasks: high quality image generation and language understanding. The generated image is required to be both visually realistic and semantically consistent to the language description. Deep learning based text-to-image synthesis can perhaps be dated back to the use of LSTM for iterative hand-writting generation [188]. This iterative image generation idea is later extended to form the deep recurrent attentive writer (DRAW) method that combines an LSTM based sequential variational auto-encoder (VAE) with a spatial attention mechanism [189]. alignDRAW modifies DRAW to use natural language based descriptions to synthesis images with general content [190]. An attention model is used to compute the alignment between the input words and the patches drawn iteratively.

1) GAN based methods: Comparing to VAE, conditional-GAN (CGAN) is found to be able to synthesis highly compelling images of specific categories that a human might mistake for real [191], [192]. A GAN model consists of a generator that synthesize candidates based on input noises and a discriminator that evaluates them. Adversarial training is employed to make the generator to capture the true data distribution so that the discriminator can no longer discriminate the synthesized data from the real ones [81]. CGAN extends the standard GAN structure by conditioning on extra category labels for both generator and discriminator. GAN-INT-CLS allows to synthesize visually plausible 64×64 images using the embeddings of natural language descriptions to replace the category labels in CGAN [193]. The automatic evaluation of the quality of text conditioned images can be less straightforward. To find the discriminability of GAN generated images, inception score (IS) [194] and Fréchet inception distance [195] (FID) are often used. Multi-scale structural similarity (MS-SSIM) [196] is commonly used to evaluate the diversity of images. To evaluate whether a generated image is semantically consistent with the input text description, R-precision [48] and visual-semantic similarity [197] are used as the metrics.

2) Generating high quality images: Though basically reflecting the meaning of the descriptions, it is found the images produced by GAN-INT-CLS do not have necessary details and vivid object parts, and therefore leads to the StackGAN method [198]. StackGAN decomposes image synthesize into more manageable sub-problems through a sketch-refinement process by stacking two CGANs trained separately. The first GAN produces 64×64 low-resolution images by sketching the primitive shape and colors of the object based on the text, and the second GAN is trained after to generate 256×256 images by rectifying defects and adding compelling details in the low-resolution image. StackGAN++ improves this idea by adding an extra GAN to generate 128×128 images in between and training all GANs jointly [199]. To ensure the generated image semantically match the text precisely, [48] proposed attentional GAN (AttnGAN), which also stacks three GANs for different image resolutions [48], and while the first GAN is conditioned on the sentence embedding, the next two GANs are conditioned on bimodal embeddings produced by attention models fusing word-level features with low-resolution images. It is shown attention mechanism can help GAN to focus on

words that are most relevant to the sub-region drawn at each stage. Apart from stacking the generators, [200] shows that high resolution images can also be generated with a dynamic memory module.

3) Generating semantically consistent images: To improve the semantic consistency between relevant image and text features, DAMSM is proposed for AttnGAN [48]. [197] tackles the same problem by leveraging hierarchical representations with extra adversarial constraints to discriminate not only real/fake image pairs, but also real/fake image-text pairs at multiple image resolutions in the discriminator, and is named as hierarchically-nested discriminator GAN (HDGAN). Similarly, text conditioned auxiliary classifier GAN (TAC-GAN) introduces an extra image classification task to the discriminator [201], whereas text-conditioned semantic classifier GAN (Text-SeGAN) alternates the classifier with a regression task to estimate the semantic relevance between image and text [202]. Analogous to cycle consistency [203], MirrorGAN is proposed to improve the semantic consistency between the two modalities using an extra image captioning module [204].

4) Semantic layout control for complex scenes: With the success in the generation of realistic and semantically consistent images for single objects, such as birds [205] or flowers [206], state-of-the-art text-to-image synthesis methods still struggle to generate complex scenes with many objects and relationships, such as those in the Microsoft COCO data set [207]. In the pioneering work [208], not only text descriptions but also locations of objects specified by keypoints or bounding boxes are used as the input. Later, detailed semantic layout, such as a scene graph, is used to replace the natural language sentence as a more direct description of objects and their relationships [209]–[211]. Meanwhile, efforts are made to keep natural language input while incorporating the idea of semantic layout. [212] includes extra object pathway to both generator and discriminator to explicit control the object locations. [213] employs a two-stage procedure that first builds a semantic layout automatically from the input sentence with LSTM based box and shape generators, and then synthesizes the image using image generator and discriminators. Since fine-grained word/object level information is not explicitly used for generation, such synthesized images do not contain enough details to make them look realistic. The object-driven attentive GAN (Obj-GAN) improves the two-stage generation idea using a pair of object-driven attentive image generator and object-wise discriminator [140]. At every generation step, the generator uses the text description as a semantic layout and synthesizes the image region within a bounding box by focusing on the words that are most relevant to the object in it. ObjGAN is found to be more robust and interpretable, and significantly improves the object generation quality for complex scenes.

5) Other topics: In addition to the layout, other types of fine-grained control in image generation have also been studied in literature. Attribute2Image [214] studies the use of attributes in face generation, such as age and gender *etc*. [215] uses the same idea for face editing, such as to remove the beard or change the hair color. Text-adaptive GAN [216] allows semantic modification of input images for birds and flowers via natural language. [217] enforces to learn the representation content and style as two disentangled variables using a dual inference mechanism based on cycle-consistency for text-to-image synthesis. The success of these methods validate GAN is able to learn some semantic concepts as disentangled representations, as in Section II-A3. Text2Scene is another interesting work that generates compositional scene representation from natural langauge step-by-step without using GANs [218]. It is shown with minor modifications, Text2Scene can generate cartoon like, semantic layout, and real image like scenes. Dialogue based interaction is studied to control image synthesis, in order to improve complex scene generation progressively [219]–[223]. Meanwhile, text-to-image synthesis is extended to multiple images or videos, where visual consistency is required among the generated images [224]–[226].

C. Visual Question Answering

1) Task definition: VQA extends text-based QA from NLP by asking questions related to the visual information presented in an image or a video clip. For image based VQA, it is often considered as a visual Turing test, in which the system is required to understand any form of natural language-based questions and to answer them in a natural way. However, it is often simplified as a classification task defined in different ways to focus on the core problem [43], [44], [132], [227], [228]. Initial works generated the questions using templates or by converting from description sentences using syntax trees [227], [229]. Later later studies focus on the use of free-form natural language questions authored either by human or powerful deep generative models, such as GAN and VAE [44], [229]–[231]. Different from the open-ended questions presented in complete sentence form, possible answers are often presented as a large set of classes (e.g. 3000) related to yes/no, counts, object classes and instances *etc*. To focus on the core understanding and reasoning problems, VQA can be simplified as to classify visual and textual features into the answer related classes.

Alternatively, VQA can be defined as to select among multiple (e.g. 4) choices, and each choice is associated with each answer presented in the form of a natural language sentence [132]. This setup can be implemented as a classification to the choices based on features of the image, question, and answer candidates [154]. There exist other types of VQA task definitions, such as the Visual Madlibs dataset that requires to answer the questions by "fill-in-the-blanks" [45]. Furthermore, visual dialogue can be viewed as to answer a sequence of questions grounded in an image [232], [233], which extends VQA by requiring to generate more human like responses and to infer the context from the dialogue history.

2) Common data sets and approaches: The first VQA data set, DAQUAR, uses real-world images with both template-based and human annotated questions [227]. COCO-QA has more QA pairs than DAQUAR by converting image descriptions from the MS COCO data set into questions [229]. Such questions are in general easier since they allow the model to rely more on the rough image rather than logical reasoning. VQA v1 and v2 are the most popular data sets for

VQA consisting of open-ended questions and both real and abstract scenes [44], [234]. A VQA Challenge based on these data sets is held annually as a CVPR workshop since 2016. Visual7W is a part of the Visual Genome data set for VQA with multiple choices [132]. It contains questions related to "what", "who", and "how" for spatial reasoning, and "where", "when", and "why" questions for high-level common-sense reasoning. The 7th type of the questions in Visual7W are the "which" questions, which are also termed as the pointing questions, whose answer choices are associated with bounding boxes of objects in the image. Approaches designed for these data sets often focus on fusing image and question vectors with the previously discussed attention- and bilinear pooling-based methods, such as SAN, co-attention, Up-Down, MCB, MLB, and BAN *etc*.

3) Integrating external knowledge source: Since most of the VQA questions in these data sets are about simple counting, colors, and object detections that do not need any external knowledge to answer, a further development of the task is to include more difficult questions that require knowing more than what the questions entail or what information is contained in the images. Both knowledge-based reasoning for VQA (KB-VQA) and fact-based VQA (FVQA) data sets incorporate structured knowledge base, which often requires extra steps to query the knowledge base that makes the method no longer an end-to-end trainable approach [235], [236]. Different from the structured knowledge bases, outside knowledge VQA (OK-VQA) uses external knowledge in the form of natural language sentences collected by retrieving Wikipedia articles with search queries extracted from the question, and an extra ArticleNet model is trained to find the answers from the retrieved articles [237].

4) Discounting language priors: Though significant achievements have been made, recent studies point out that the common VQA benchmarks suffer from strong and prevalent priors – most bananas are yellow and mostly the sky is blue, which can often cause the VQA model to over-fit to these statistical biases and tendencies from the answer distributions, and largely circumvent the need to understand the visual scenes. Based on the objects, attributes, and relations provided through the scene graphs from Visual Genome, a new data set, GQA, was created to greatly reduce such biases by generating questions with a functional program that controls the reasoning steps [238]. New splits for VQA v1 and VQA v2 are generated to have different answer distributions for every question of the training and test sets, which are referred to as VQA under challenging priors (VQA-CP v1 and VQA-CP v2) [239]. Recent methods propose to handle the biased priors with adversarial training or additional train only structures [240], [241].

5) Other issues: Another problem that current VQA methods suffers from is the low robustness against linguistic variations from the questions. A data set, VQA-Rephrasings, modified the VQA v2.0 validation set with human authored rephrasing of the questions [203]. A cycle-consistency [242] based method that improves the linguistic robustness by enforcing consistencies between the original and rephrased questions, and between the true answer and the answers predicted based on the original and rephrased questions. [243] suggests that attention mechanism can cause VQA models to suffer from counting the object proposals, and an extra model component was proposed as a solution. Moreover, it is the fact that the current VQA methods cannot even read text from images. A method is proposed to address this problem by fusing not text extracted from the image using optical character recognition [244]. VizWiz is a goal oriented VQA data set collected by blind people taking possibly low quality pictures and asking questions in spoken English, which also include many text related questions [245].

D. Visual Reasoning

This section focuses on the study of a very interesting problem – visual reasoning, which is about how to conduct accurate, explicit, and expressive understanding and reasoning. Visual reasoning can involved by many language and vision based bimodal tasks, such as caption generation and text-to-image synthesis. However, in this section we mostly focus on the related methods for VQA as visual reasoning is particularly important when answering complicated questions. SAN is often considered as a pioneering work related to implicit visual reasoning since to its stacked structure can be viewed as to perform multiple reasoning steps. Shortly afterwards, feature-wise linear modulation (FiLM) is proposed to refine visual features iteratively using feature-wise affine transforms based on the scaling factors and bias values generated dynamically from the textual features [246]. Multimodal relational network (MuRel) also has a structure with multiple MuRel cells based on bilinear pooling, which can be used iteratively [247].

1) Neural module network based methods: Neural module network (NMN) is a method which composes a collection of jointly trained neural modules into a deep model for answering the question [248]. A dependency parser first helps to convert the natural language question into a fixed and rule-based network layout, and specify both the set of modules used to answer the question and the connections between them. Then a deep model is assembled based on the layout to produce the prediction of the answers. SHAPES, a synthetic dataset consists of complex questions about simple arranges of ordered shapes, was also proposed to focus on the compositional phenomena of questions [248]. A later study learns the model layout predictor jointly with the parameters of the modules by re-ranking a list of layout candidates using reinforcement learning, which is termed as dynamic NMN (D-NMN) [249]. Modules such as "find" or "relate" operation uses attention models to focus on one or two regions of the input image and makes the forwarding of the assembled deep model similar to running a functional program [249]. An end-to-end version of NMN (N2NMN) used an RNN question encoder to convert the input question into a layout policy without requiring the aid of a parser [250]. The work is based on a more recent data set called compositional language and elementary visual reasoning diagnostics (CLEVR). As its name suggests, CLEVR is a synthetic diagnostic data set testing a range of visual reasoning abilities of objections and relationships with minimal biases and detailed annotations describing the kind of

reasoning each question requires [251]. Other implementations of NMN include the program generator and execution engine method (PG+EE) that shares generic design among some operations [252], stack-NMN that improves the parser and incorporates question features into the modules [253], and transparency by design network (TbD-net) that redesigns some modules from PG+EE to maintain the transparency of the reasoning procedure [254].

2) Other types of end-to-end reasoning methods: Another end-to-end approach is the memory, attention, and composition (MAC) network that decomposes the question into a series of attention-based reasoning steps and perform each of them using a recurrent MAC cell that maintains a separation between the control and memory hidden states. Each hidden state is generated by an ANN model constructed based on attention and gating mechanisms [255]. More recently, both deterministic symbolic programs and probabilistic symbolic models have been used as the execution engine for the generated programs to improve the transparency and data efficiency, which result in the neural-symbolic VQA (NS-VQA) and probabilistic neural-symbolic models (prob-NMN) respectively [256], [257]. As an extension of NS-VQA, the neuro-symbolic concept learner (NS-CL) uses a neuro-symbolic reasoning module to execute programs on the scene representation. NS-CL can have its program generator, reasoning module, and visual perception components jointly trained in an end-to-end fashion without requiring any component-level supervisions [258]. Its perception module learns visual concepts based on the language descriptions of the objects and facilitates learning new words and parsing new sentences.

We finish this section by reviewing the relation networks (RN), which has a simple structure that uses an ANN as the function to model the relationship between any pair of visual and textual features, and the resulted output values are accumulated and transformed by another ANN [259]. Though RN merely models the relationship without any form of induction reasoning, it achieves very high VQA accuracy on CLEVR. This inspires a re-thinking of the connection between correlation and induction.

V. Conclusion

This paper reviews the area of modeling and machine learning across multiple modalities based on deep learning, particularly the combination of vision and natural language. In particular, we propose to organize the many pieces of work in the language-vision multimodal intelligence field from three aspects, which include multimodal representations, the fusion of multimodal signals, and the applications of multimodal intelligence. In the section of representations, both single modal and multimodal representations are reviewed under the key concept of embedding. The multimodal representation unifies the involved signals of different modalities into the same vector space for general downstream tasks. On multimodal fusion, special architectures, such as attention mechanism and bilinear pooling, are discussed. In the application section, three selected areas of broad interest are presented, which include image caption generation, text-to-image synthesis, and visual question answering. A set of visual reasoning methods for VQA is also discussed. Our review covers task definition, data set specification, development of commonly used methods, as well as issues and trends, and therefore can facilitate future studies in this emerging field of multimodal intelligence for our community.

References

[1] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[2] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, pp. 1–127, 2009.

[3] L. Deng and Y. Dong, "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing*, vol. 7, pp. 197–387, 2014.

[4] J. Schmidhuber, "Deep learning in neural networks: An overview," *NeuralNetworks*, 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 Learning*. The 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," *Proc. NIPS Workshop*, 2010.

[8] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep autoencoder," *Proc. Interspeech*, 2010.

[9] L. Deng, "An overview of deep-structured learning for information processing," in *Proc. APSIPA ASC*, 2011.

[10] D. Yu, L. Deng, F. Seide, and G. Li, "Discriminative pre-training of deep nerual networks," in *U.S. Patent No. 9,235,799*, 2011.

[11] G. Dahl, D. Yu, and L. Deng, "Large-vocabulry continuous speech recognition with context-dependent DBN-HMMs," in *Proc. ICASSP*, 2011.

[12] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. ICASSP*, 2013.

[13] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 30–42, 2012.

[14] F. Seide, L. Gang, and Y. Dong, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, 2011.

[15] G. Hinton, L. Deng, Y. Dong, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[16] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," *Proc. ICASSP*, 2013.

[17] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2014.

[24] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015.

[26] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 530–539, 2015.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[28] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014.

[29] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," in *arXiv:1609.08144*, 2016.

[30] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, 2015.

[31] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. NAACL*, 2018.

[32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-yraining," in *https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/language understanding paper.pdf*, 2018.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[34] H.-Y. Shum, X. He, and D. Li, "From Eliza to XiaoIce: Challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, 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. Chapter 14 in Book: The Handbook of Multimodal-Multisensor Interfaces*. ACM and Morgan & Claypool Publishers, 2019.

[36] L. Deng and Y. Liu, *Deep Learning in Natural Language Processing*. Springer, 2018.

[37] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proc. EMNLP*, 2014.

[38] L. Yu, P. Poirson, S. Yang, A. Berg, and T. Berg, "Modeling context in referring expressions," in *Proc. ECCV*, 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," in *Proc. ICCV*, 2015.

[40] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. CVPR*, 2015.

[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015.

[42] J. Johnson, A. Karpathy, and F.-F. Li, "Densecap: Fully convolutional localization networks for dense captioning," in *Proc. CVPR*, 2016.

[43] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual Turing test for computer vision systems," in *Proc. NAS*, 2015.

[44] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batral, C. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. ICCV*, 2015.

[45] L. Yu, E. Park, A. Berg, and T. Berg, "Visual Madlibs: Fill in the blank description generation and question answering," in *Proc. ICCV*, 2015.

[46] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Proc. ECCV*, 2016.

[47] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. ICML*, 2016.

[48] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. CVPR*, 2018.

[49] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. CVPR*, 2018.

[50] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, pp. 141–151, 2000.

[51] M. Cookea, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of Acoustic Society of America*, vol. 120, pp. 2421–2424, 2006.

[52] T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. Early Access, pp. 1–13, 2018.

[53] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for video broadcast news: Some fusion techniques," in *Proc. MMSP*, 1999.

[54] 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.), pp. 493–499, Springer Berlin Heidelberg, 2005.

[55] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[56] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1086–1099, 2018.

[57] J. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," in *Proc. Interspeech*, 2019.

[58] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. ASRU*, 2019.

[59] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, pp. 112:1–11, 2018.

[60] T. Afouras, J. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018.

[61] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013.

[62] P.-S. Huang, X. He, G. J., L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. CIKM*, 2013.

[63] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proc. WWW*, 2014.

[64] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, pp. 694–707, 2016.

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

[66] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[67] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013.

[68] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*, 2013.

[69] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[70] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[71] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F.-F. Li, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

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

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

[74] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014.

[75] A. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," in *Proc. WWW*, 2015.

[76] 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. NAACL*, 2015.

[77] W.-T. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proc. ACL*, 2014.

[78] 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. ACL*, 2015.

[79] R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," in *Proc. NIPS*, 2015.

[80] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. NAACL HLT*, 2013.

[81] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014.

[82] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016.

[83] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011.

[84] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. NIPS*, 2012.

[85] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. ACL*, 2014.

[86] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, *et al.*, "From captions to visual concepts and back," in *Proc. CVPR*, 2015.

[87] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional semantics in technicolor," in *Proc. ACL*, 2012.

[88] S. Kottur, R. Vedantam, J. Moura, and D. Parikh, "Visual Word2Vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes," in *Proc. CVPR*, 2016.

[89] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in *Proc. CVPR*, 2017.

[90] P. Bachman, R. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. NeurIPS*, 2019.

[91] A. Lazaridou, N. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," in *Proc. NAACL*, 2015.

[92] A. Karpathy, A. Joulin, and F.-F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. NIPS*, 2014.

[93] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W.-Y. Ma, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *Proc. CVPR*, 2019.

[94] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. ECCV*, 2018.

[95] Y.-H. Tsai, P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. ICLR*, 2018.

[96] T. Gupta, A. Schwing, and D. Hoiem, "ViCo: Word embeddings from visual co-occurrences," in *Proc. ICCV*, 2019.

[97] D.-K. Nguyen and T. Okatani, "Multi-task learning of hierarchical vision-language representation," in *Proc. CVPR*, 2019.

[98] R. Socher, M. Ganjoo, H. Sridhar, M. C. Bastani, O., and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. NIPS*, 2013.

[99] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. NIPS*, 2013.

[100] Y.-H. Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proc. ICCV*, 2017.

[101] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proc. ICCV*, 2015.

[102] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," in *Proc. CVPR*, 2016.

[103] G. Collell and M.-F. Moens, "Do neural network cross-modal mappings really bridge modalities?," in *Proc. ACL*, 2018.

[104] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[105] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 37, pp. 328–339, 1989.

[106] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," in *Proc. ICLR*, 2017.

[107] 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," in *arXiv:1908.06066*, 2019.

[108] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visuallinguistic representations," in *arXiv:1908.08530*, 2019.

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

[110] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. ICCV*, 2019.

[111] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," in *Proc. ICMLC*, 2019.

[112] H. Tan and B. Mohit, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. EMNLP*, 2019.

[113] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. NeurIPS*, 2019.

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

[115] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. ACL*, 2019.

[116] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *Proc. CVPR*, 2012.

[117] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NIPS*, 2014.

[118] T. Wörtwein and S. Scherer, "What really matters An information gain analysis of questions and reactions in automated PTSD screenings," in *Proc. ACII*, 2017.

[119] G. Ye, D. Liu, I.-H. Jhuo, and C. S.-F., "Robust late fusion with rank minimization," in *Proc. CVPR*, 2012.

[120] B. Nojavanasghari, D. Gopinath, J. Koushik, B. T., and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. ICMI*, 2016.

[121] H. Wang, A. Meghawat, L.-P. Morency, and E. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. ICME*, 2017.

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

[123] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Proc. ECCV*, 2018.

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

[125] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. CVPR*, 2019.

[126] B. Zoph and Q. Le, "Neural architecture search with reinforcement learning," in *Proc. ICLR*, 2017.

[127] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, F.-F. Li, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proc. ECCV*, 2018.

[128] J.-M. Pérez-Rúa, M. Baccouche, and S. Pateux, "Efficient progressive neural architecture search," in *Proc. BMVC*, 2019.

[129] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proc. ACM MM*, 2016.

[130] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

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

[132] Y. Zhu, O. Groth, M. Bernstein, and F.-F. Li, "Visual7W: Grounded question answering in images," in *Proc. CVPR*, 2016.

[133] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015.

[134] K. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. CVPR*, 2016.

[135] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. CVPR*, 2016.

[136] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. ECCV*, 2016.

[137] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. ICML*, 2016.

[138] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. CVPR*, 2018.

[139] 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," in *Proc. AAAI*, 2018.

[140] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in *Proc. CVPR*, 2019.

[141] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. NIPS*, 2016.

[142] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. CVPR*, 2017.

[143] H. Fan and J. Zhou, "Stacked latent attention for multimodal reasoning," in *Proc. CVPR*, 2018.

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

[145] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," in *Proc. NIPS*, 2017.

[146] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. González, "Gated multimodal units for information fusion," in *Proc. ICLR*, 2017.

[147] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual QA," in *Proc. NIPS*, 2016.

[148] H. Noh, P. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. CVPR*, 2016.

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

[150] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*, 2017.

[151] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. CVPR*, 2016.

[152] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Proc. ICALP*, 2012.

[153] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. SIGKDD*, 2013.

[154] 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. EMNLP*, 2016.

[155] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. ICLR*, 2017.

[156] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. ICCV*, 2017.

[157] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 5947–5959, 2018.

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

[159] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. ICCV*, 2017.

[160] L. Lathauwer, "Decompositions of a higher-order tensor in block termspart II: Definitions and uniqueness," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, pp. 1033–1066, 2008.

[161] H. Ben-younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proc. AAAI*, 2019.

[162] Z. Liu, Y. Shen, V. Lakshminarasimhan, P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. ACL*, 2018.

[163] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. NeurIPS*, 2018.

[164] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. CVPR*, 2018.

[165] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris, "Dialog-based interactive image retrieval," in *Proc. CVPR*, 2018.

[166] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proc. CVPR*, 2018.

[167] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, and N. Sünderhauf, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. CVPR*, 2018.

[168] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge, "Stay on the path: Instruction fidelity in vision-and-language navigation," in *Proc. ACL*, 2019.

[169] R. Hu, D. Fried, A. Rohrbach, D. Klein, T. Darrell, and K. Saenko, "Are you looking? Grounding to multiple modalities in vision-and-language navigation," in *Proc. ACL*, 2019.

[170] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments," in *Proc. CVPR*, 2019.

[171] L. Ke, X. Li, Y. Bisk, A. Holtzman, Z. Gan, J. Liu, J. Gao, Y. Choi, and S. Srinivasa, "Tactical rewind: Self-correction via backtracking in vision-and-language navigation," in *Proc. CVPR*, 2019.

[172] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proc. CVPR*, 2019.

[173] C.-Y. Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, and C. Xiong, "Self-monitoring navigation agent via auxiliary progress estimation," in *Proc. ICLR*, 2019.

[174] J. Fu, A. Korattikara, S. Levine, and S. Guadarrama, "From language to goals: Inverse reinforcement learning for vision-based instruction following," in *Proc. ICLR*, 2019.

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

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

[177] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015.

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

[179] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. CVPR*, 2015.

[180] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015.

[181] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. CVPR*, 2017.

[182] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proc. CVPR*, 2017.

[183] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *Proc. CVPR*, 2019.

[184] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*, 2016.

[185] K. Tran, X. He, L. Zhang, and J. Sun, "Rich image captioning in the wild," in *Proc. CVPR Workshop*, 2016.

[186] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. CVPR*, 2017.

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

[188] A. Graves, "Generating sequences with recurrent neural networks," in *arXiv:1308.0850*, 2013.

[189] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proc. ICML*, 2015.

[190] E. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov, "Generating images from captions with attention," in *Proc. ICLR*, 2016.

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

[192] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Proc. NIPS*, 2015.

[193] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. ICML*, 2016.

[194] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, 2016.

[195] 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. NIPS*, 2017.

[196] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," 2017.

[197] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. CVPR*, 2018.

[198] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. ICCV*, 2017.
[199] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1947–1962, 2019.
[200] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. CVPR*, 2019.
[201] A. Dash, J. Gamboal, S. Ahmed, M. Liwicki, and M. Afzal, "TAC-GAN – Text conditioned auxiliary classifier generative adversarial network," in *Proc. CVPR*, 2017.
[202] M. Cha, Y. Gwon, and H. Kung, "Adversarial learning of semantic relevance in text to image synthesis," in *Proc. AAAI*, 2019.
[203] X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," in *Proc. CVPR*, 2019.
[204] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. CVPR*, 2019.
[205] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.
[206] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. CVPR*, 2006.
[207] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014.
[208] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. NIPS*, 2016.
[209] J. Johnson, A. Gupta, and F.-F. Li, "Image generation from scene graphs," in *Proc. CVPR*, 2018.
[210] S. Tripathi, A. Bhiwandiwalla, A. Bastidas, and H. Tang, "Heuristics for image generation from scene graphs," in *Proc. ICLR Workshop LLD*, 2019.
[211] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," in *Proc. CVPR*, 2019.
[212] T. Hinz, S. Heinrich, and S. Wermter, "Generating multiple objects at spatially distinct locations," in *Proc. ICLR*, 2019.
[213] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. CVPR*, 2018.
[214] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Proc. ECCV*, 2016.
[215] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, pp. 5464–5478, 2019.
[216] S. Nam, Y. Kim, and S. Kim, "Text-adaptive generative adversarial networks: Manipulating images with natural language," in *Proc. NeurIPS*, 2018.
[217] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L. Jorio, and T. Fevens, "Dual adversarial inference for text-to-image synthesis," in *Proc. ICCV*, 2019.
[218] F. Tan, S. Feng, and V. Ordonez, "Text2Scene: Generating compositional scenes from textual descriptions," in *Proc. CVPR*, 2019.
[219] S. Sharma, D. Suhubdy, V. Michalski, S. Kahou, and Y. Bengio, "ChatPainter: Improving text to image generation using dialogue," in *Proc. ICLR Workshop*, 2018.
[220] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. Asri, S. Kahou, Y. Bengio, and G. Taylor, "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction," in *Proc. ICCV*, 2019.
[221] P. Cascante-Bonilla, X. Yin, V. Ordonez, and S. Feng, "Chat-crowd: A dialog-based platform for visual layout composition," in *Proc. NAACL-HLT*, 2018.
[222] Y. Chen, Z. Gan, Y. Li, J. Liu, and J. Gao, "Sequential attention GAN for interactive image editing via dialogue," in *Proc. AAAI*, 2019.
[223] J.-H. Kim, N. Kitaev, X. Chen, M. Rohrbach, B.-T. Zhang, Y. Tian, D. Batra, and D. Parikh, "CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication," in *Proc. ACL*, 2019.
[224] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, "StoryGAN: A sequential conditional GAN for story visualization," in *Proc. CVPR*, 2019.
[225] Y. Li, M. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text," in *Proc. AAAI*, 2018.
[226] Y. Balaji, M. Min, B. Bai, R. Chellappa, and H. Graf, "Conditional GAN with discriminative filter generation for text-to-video synthesis," in *Proc. IJCAI*, 2019.
[227] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on yncertain input," in *Proc. NIPS*, 2014.
[228] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. ICCV*, 2015.
[229] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. NIPS*, 2015.
[230] 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," in *Proc. CVPR*, 2018.
[231] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," in *Proc. CVPR*, 2017.
[232] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. Moura, D. Parikh, and D. Batra, "Visual dialogue," in *Proc. CVPR*, 2017.
[233] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "GuessWhat?! Visual object discovery through multi-modal dialogue," in *Proc. CVPR*, 2017.
[234] 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," *International Journal of Computer Vision*, vol. 127, pp. 398–414, 2019.
[235] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Explicit knowledge-based reasoning for visual question answering," in *Proc. IJCAI*, 2017.
[236] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, "FVQA: Fact-based visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 2413–2427, 2018.
[237] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *Proc. CVPR*, 2019.
[238] D. Hudson and C. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. CVPR*, 2019.
[239] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Dont just assume; Look and answer: Overcoming priors for visual question answering," in *Proc. CVPR*, 2018.
[240] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. NeurIPS*, 2018.
[241] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases in visual question answering," in *Proc. NeurIPS*, 2019.
[242] J.-Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017.
[243] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," in *Proc. ICLR*, 2018.
[244] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proc. CVPR*, 2019.
[245] D. Gurari, Q. Li, A. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. Bigham, "VizWiz grand challenge: Answering visual questions from blind people," in *Proc. CVPR*, 2018.
[246] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018.
[247] R. Cadene, H. Ben-younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," in *Proc. CVPR*, 2019.
[248] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," in *Proc. CVPR*, 2016.
[249] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proc. NAACL*, 2016.
[250] 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. ICCV*, 2017.
[251] 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," in *Proc. CVPR*, 2017.
[252] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F.-F. Li, C. Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *Proc. ICCV*, 2017.
[253] R. Hu, J. Andreas, T. Darrell, and K. Saenko, "Explainable neural computation via stack neural module networks," in *Proc. ECCV*, 2018.

[254] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proc. CVPR*, 2018.
[255] D. Hudson and C. Manning, "Compositional attention networks for machine reasoning," in *Proc. ICLR*, 2018.
[256] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," in *Proc. NeurIPS*, 2018.
[257] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural-symbolic models for interpretable visual question answering," in *Proc. ICML*, 2018.
[258] J. Mao, C. Gan, P. Kohli, J. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *Proc. ICLR*, 2019.
[259] A. Santoro, D. Raposo, D. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Proc. NIPS*, 2017.

Chao Zhang is an advisor of JD.com speech team, and a research associate in speech and natural language processing at the University of Cambridge. He received his B.E. and M.S. degrees in 2009 and 2012 respectively, both from the Department of Computer Science & Technology, Tsinghua University, and a Ph.D. degree in 2017 from Cambridge University Engineering Department.

Zichao Yang is a quantitative researcher at Citadel. Prior to that, he received his Phd in computer science from Carnegie Mellon University. His research interests are in machine learning, deep learning and their applications in natural language processing and computer vision. He has published dozens of papers in NeurIPS, ICML, CVPR, ICCV, EMNLP, NAACL *etc*.

Xiaodong He (IEEE Member 2003, Senior member 2008, Fellow 2019) is the Deputy Managing Director of JD AI Research, and Head of the Deep learning, NLP and Speech Lab. He is also Affiliate Professor of ECE at the University of Washington (Seattle). His research interests are mainly in deep learning, natural language processing, speech recognition, computer vision, information retrieval, and multimodal intelligence. He has held editorial positions on multiple IEEE Journals and the Transactions of the ACL, and served in the organizing committee/program committee of major speech and language processing conferences. He is a member of the IEEE SLTC for the term of 2015-2017 and the Chair of the IEEE Seattle Section in 2016. He received the Bachelor degree from Tsinghua University in 1996, MS degree from Chinese Academy of Sciences in 1999, and the PhD degree from the University of Missouri – Columbia in 2003.

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

He was an elected member of the Board of Governors of the IEEE Signal Processing Society, and was Editors-in-Chief of IEEE Signal Processing Magazine and of IEEE/ACM Transactions on Audio, Speech, and Language Processing (2008-2014), for which he received the IEEE SPS Meritorious Service Award. In recognition of the pioneering work on disrupting 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 the contributions to artificial intelligence, machine learning, information retrieval, multimedia signal processing, speech processing and recognition, and human language technology. He is an author or co-author of six technical books on deep learning, speech processing, pattern recognition and machine learning, and, the latest, natural language processing (Springer, June 2018).