Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

Aditya Mogadala amogadala@lsv.uni-saarland.de
Marimuthu Kalimuthu mkalimuthu@lsv.uni-saarland.de
Dietrich Klakow dietrich.klakow@lsv.uni-saarland.de
Spoken Language Systems (LSV), Saarland University,
Saarland Informatics Campus, 66123 Saarbrücken, Germany

Abstract

Integration of vision and language tasks has seen a significant growth in the recent times due to surge of interest from multi-disciplinary communities such as deep learning, computer vision, and natural language processing. In this survey, we focus on ten different vision and language integration tasks in terms of their problem formulation, methods, existing datasets, evaluation measures, and comparison of results achieved with the corresponding state-of-the-art methods. This goes beyond earlier surveys which are either task-specific or concentrate only on one type of visual content i.e., image or video. We then conclude the survey by discussing some possible future directions for integration of vision and language research.

1. Introduction

Recent advancements in deep learning research has lead the fields of computer vision (CV) and natural language processing (NLP) see a significant progress in several tasks. Independent from NLP, CV has achieved prominent improvements in tasks such as visual content classification (He et al., 2016), object detection (Redmon & Farhadi, 2017) and segmentation (He et al., 2017) etc., using large annotated datasets or self-supervision (Jing & Tian, 2019). Similarly, independent from CV, NLP has seen surge of interest in solving multiple tasks at once with unsupervised pretraining of language models (Devlin et al., 2018; Radford et al., 2019; Lample & Conneau, 2019) using large unlabeled corpora. However, there is also interest in solving challenges that combine linguistic and visual information from these traditionally independent fields. The methods which address the challenge of integration are supposed to provide complete understanding of visual or textual content and are expected to (1) Generate comprehensible but concise and grammatically well-formed descriptions about the visual content or vice versa where given a textual description generate the visual content back (2) Identify the objects in the visual content and infer their relationships to reason or answer arbitrary questions about them (3) Navigate through an environment by leveraging input from both vision and natural language instructions (4) Translate textual content from one language to another with visual content used for disambiguation (5) Generate stories about the visual content and so on. Design of these methods which can process and relate information from multiple modalities (i.e., linguistic and visual information) are usually referred to be sub-part of multimodal learning models (Mogadala, 2015).

Efficiently solving challenges mentioned above can result in many potential applications. For example, it can assist visually impaired individuals to get a holistic visual scene under-

standing where a person can get information about a scene from its generated descriptions, stories and further asking arbitrary questions about it. Other applications include automatic surveillance (Baumann et al., 2008), autonomous driving (Kim et al., 2018), human-computer interaction (Rickert et al., 2007), city navigation (de Vries et al., 2018) and so on. Also, solving such challenges can provide an excellent test bed for CV and NLP systems, one that is much more comprehensive than independent CV and NLP evaluations.

Given such a broad scope for fundamental and applied research, there are several surveys proposed in the recent years to provide comprehensive overview of the integration of vision and language tasks. These surveys have concentrated on covering specific language and vision integration task such as image (Bernardi et al., 2016; Bai & An, 2018; Hossain et al., 2019) or video (Aafaq et al., 2018) description generation, visual question answering (Kafle & Kanan, 2017; Wu et al., 2017b), action recognition (Gella & Keller, 2017) and visual semantics (Liu et al., 2019). Surveys which went beyond a specific task have summarized dataset statistics (Ferraro et al., 2015), provided comprehensive overview of only natural language processing tasks such as natural language generation (NLG) (Gatt & Krahmer, 2018) and commonsense reasoning (Storks et al., 2019). However, there are attempts to cover multiple modalities (including sound) (Baltrušaitis et al., 2019), but is structured in a bottom-up manner giving more importance to underlying fusion technologies than the task. Also, there is an interest to understand limitations (Kafle et al., 2019) in integration of vision and language research. However, it is only limited to language grounded image understanding tasks. There was also interest from a theoretical point of view to develop theories on the complementarity of language and visual data in the human-machine communication (Moens et al., 2019).

Nevertheless, in this survey, we go beyond and provide comprehensive overview of ten different tasks along with methods, datasets, and evaluation metrics that are driving the current integration of vision and language research. We first introduce ten prominent tasks of integration of vision and language along with their methods in the Section 2 and give an overview of the datasets used for each of these tasks in Section 3. Then, in Section 4, we describe the representation used for both vision and language separately and further discuss the major methods that combine both of them for realizing a task. In Section 5, we present the evaluation metrics used for all ten tasks. Furthermore, in Section 6, benchmark results achieved for each task through the corresponding methods are compared and discussed. In Section 7, we discuss the possible future directions and finally Section 8 concludes our survey and discusses some insights into the findings.

2. Tasks

Over the past few years significant research is performed in integrating language and vision. Several tasks are proposed which combine language observed at different levels such as words, phrases, sentences, paragraphs and documents with visual information represented with images or videos. Initially, most of the work was concentrated on combining low-level linguistic units such as words with images or videos for building visual-semantic embeddings (Frome et al., 2013; Kiros et al., 2014b; Liu et al., 2015; Cao et al., 2016; Tsai et al., 2017; Guo et al., 2018; Wang et al., 2019) useful for further down stream applications.

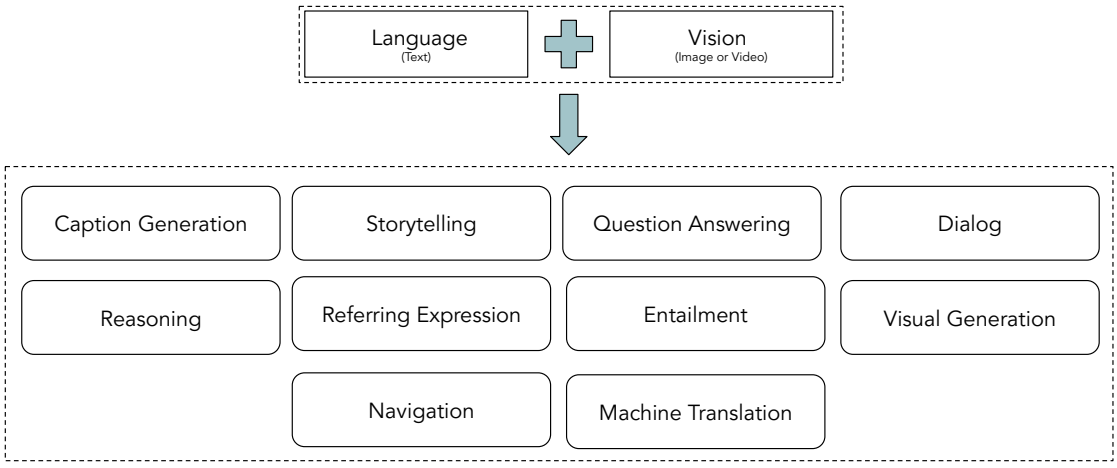

Figure 1: Ten different Language and Vision integration tasks.

In our survey, however, we go beyond words and present those tasks which consider variable-length text larger than words as the language input. Most of these tasks are seen as an extension to either CV, NLP or both problems. Figure 1 summarizes different tasks. However, to get a feel about how these tasks are seen as a natural extension of problems in CV, NLP or both, we briefly find the correlation of them with similar tasks addressed in their individual research.

Extension of NLP Tasks

- **Visual Description Generation** is closely aligned with conditional language modeling (De Mulder et al., 2015) or as a natural language generation (NLG) (Reiter & Dale, 2000) challenge in NLP. Given a non-linguistic information (e.g., image or video), the goal is to generate a human-readable text snippet that describes the input.
- The task of **Visual Storytelling** forms similar hypothesis as visual description generation. However, instead of dealing with single visual input, sequences of visual inputs are used to generate a narrative summary based on the aligned text with them. It can be seen that the task closely aligned to text summarization (Nallapati et al., 2016; Liu et al., 2018) mostly generating abstractive summaries.
- **Visual Question Answering** draws its inspiration from the text-based question-answering (Harabagiu et al., 2000; Strzalkowski & Harabagiu, 2006) which is one of the long pursued NLP research topic. Here, answering questions about visual information is seen as its natural extension.
- The task of **Visual Dialog** aims at a meaningful dialog in natural, conversational language about visual content. It is seen as an visual analogue of the text-based dialog and conversation system (Weizenbaum, 1966; Dodge et al., 2015; Li et al., 2016) explored in NLP over many years.
- **Visual Referring Expression** is seen as extension of referring expression (Krahmer & Van Deemter, 2012) in natural language generation systems. Also, sub-problem

in visual referring expression i.e., comprehension is seen as analogy of pragmatics in linguistics (Thomas, 2014) due to its use of context.

- **Visual Entailment** is seen as an inference task for predicting whether the image semantically entails the text. It is seen as a natural extension to natural language inference (Condoravdi et al., 2003; Bowman et al., 2015), where the premise is text, instead of a visual premise.

- **Multimodal Machine Translation** aims to perform translation from source to target language by leveraging visual information along with the source language. It is been influenced from the well known NLP challenge of automatic translation between two languages (Brown et al., 1990; Bahdanau et al., 2014).

Extension of CV Tasks

- **Visual Generation** deals with the generation of visual content by conditioning on text. It can be seen as a multimodal extension of the popular CV tasks of image-to-image translation (Isola et al., 2017) and style transfer (Gatys et al., 2016).

- The task of **Visual Reasoning** is perceived as a direct extension of the visual perception where standard CV tasks such as object classification (Krizhevsky et al., 2012), detection (Ren et al., 2015b) or segmentation (Long et al., 2015) are performed. Instead of providing only class labels (in case of classification), bounding boxes (in case of detection) or segments (in case of segmentation), visual reasoning is expected to provide relationship between detected objects by generating an entire visual scene graph.

Extension of both NLP and CV Tasks

- **Vision-and-Language Navigation** is one such task which can be seen as a transition from standard vision based navigation using only visual input (Sinopoli et al., 2001; Blösch et al., 2010) or natural language instruction based navigation (MacMahon et al., 2006; Vogel & Jurafsky, 2010). The expectation here is that, a natural language navigation instruction should be interpreted on the basis of what is visually seen. Hence, it carries out a process of combining both vision and language.

In the following, we will cover each of these tasks in detail along with the methods proposed to address them.

2.1 Visual Description Generation

Description generation aims to generate either global or dense descriptions to a given visual input. However, there can be various ways to explore the problem with different types of visual input i.e., either image or a video.

2.1.1 Image Description Generation

Standard Image Description Generation The goal of the standard image description generation is to generate sentence-level descriptions given an image. They leverage the

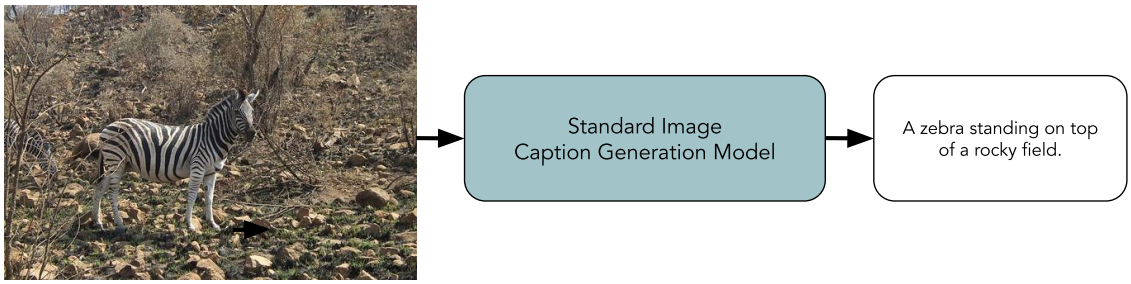

Figure 2: Given an image, the standard image description generation model generates a global textual description.

vocabulary of the dataset to generate the best description that depicts the image. Figure 2 summarizes the task.

Initially, several methods are proposed based on templates, n-grams and dependency parsing (Farhadi et al., 2010; Yang et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kulkarni et al., 2013; Elliott & Keller, 2013; Fang et al., 2015). Recently, however encoder-decoder framework (Cho et al., 2014) based image description generation models became popular and have been extended with attention mechanism (Bahdanau et al., 2014) to support the selection of local image features that are useful for generation of the word at each time step. Table 1 summarizes different setups for generating image descriptions using neural network based non-attention, attention and reinforcement learning (RL) approaches. Other variations include cross-lingual image captioning (Miyazaki & Shimizu, 2016) and multi-language image description generation (Elliott et al., 2015).

In the following, we explore some of the related ideas which expand the scope of image description generation.

Dense Image Description Generation aims to generate descriptions at the local object-level and are referred as dense captions. Several approaches (Plummer et al., 2015; Johnson et al., 2016; Rohrbach et al., 2016a; Hu et al., 2017a) are proposed to generate dense captions. Usually, they represent phrases and also their relationships to generate descriptions (Kim et al., 2019).

Image Paragraph Generation generates paragraphs instead of generating a dense or single description for an image. Generated paragraphs are expected to be coherent and fine-grained natural language descriptions (Krause et al., 2017; Liang et al., 2017).

Spoken Language Image Description Generation expands image description generation that work only with the written language to spoken language. Approaches such as visually grounded speech signal (Chrupała et al., 2017) addressed the standard image description generation from the perspective of spoken language.

Stylistic Image Description Generation adds styles to the standard image description generation where generated descriptions adhere to a specific style. For example, (Mathews et al., 2016) generated captions which captures sentiment from the image. While, (Gan et al., 2017) have generated humorous and romantic captions. It is also extended by lever-

Approach	Attention	RL
MLBL (Kiros et al., 2014a)	✗	✗
m-RNN (Mao et al., 2014)	✗	✗
Minds Eye (Chen & Lawrence Zitnick, 2015)	✗	✗
BRNN (Karpathy & Fei-Fei, 2015)	✗	✗
NIC (Vinyals et al., 2015)	✗	✗
LRCN (Donahue et al., 2015)	✗	✗
Guided LSTM (Jia et al., 2015)	✗	✗
Deep Bidirectional LSTM (Wang et al., 2016)	✗	✗
Regional Visual Attributes (Wu et al., 2017a)	✗	✗
Language CNN (Gu et al., 2017)	✗	✗
ConceptNet-NIC (Zhou et al., 2019)	✗	✗
Visual Attention (Xu et al., 2015a)	✓	✗
Region-based Attention (Jin et al., 2015)	✓	✗
Attribute Attention (You et al., 2016)	✓	✗
Review Attention (Yang et al., 2016)	✓	✗
Adaptive Attention (Lu et al., 2016)	✓	✗
Areas of Attention (Pedersoli et al., 2016)	✓	✗
Contrastive Adaptive Attention (Dai & Lin, 2017)	✓	✗
Neural Baby Talk w/ Attention (Lu et al., 2018)	✓	✗
Convolutional Attention (Aneja et al., 2018)	✓	✗
Self-Critical Attention (Rennie et al., 2016)	✓	✓
Policy Gradient (Liu et al., 2017)	✓	✓
Up-Down (Anderson et al., 2017b)	✓	✓
Multi-task Captioning (Zhao et al., 2018)	✓	✓
Stack Captioning (Gu et al., 2018)	✓	✓

Table 1: Summary of methods generating global description for an image.

aging the unpaired textual corpora (Mathews et al., 2018) to generate a story like captions. Furthermore, to make the generated captions more human like, personality traits (Shuster et al., 2018) where explored to generate captions. Recently, multi-style image description generation (Guo et al., 2019) is explored where the single model using unpaired data is built to generate different stylized captions.

Unseen Objects Image Description Generation leverage those images which lack paired descriptions. Most of the paired image-description datasets have limited visual objects to represent. Hence, methods such as Deep Compositional Captioning (DCC) (Hendricks et al., 2016), Novel object Captioner (NOC) (Venugopalan et al., 2017), Constrained Beam Search (CBS) (Anderson et al., 2017a) and LSTM-C (Yao et al., 2017) address the challenge of generating description for these images. They generate descriptions for visual object categories previously unseen in image-description corpora either by transferring information between seen and unseen objects before inference (i.e., before test time) or by keeping constraints on the generation of description words during inference (i.e., during test time). Few approaches (Mogadala et al., 2018; Lu et al., 2018) have transferred information

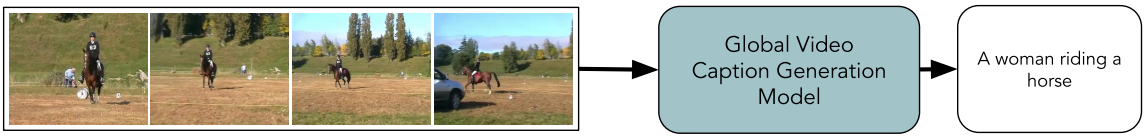

Figure 3: Given a video (represented as sequence of frames), the video description generation model generates a single global caption.

both before and during inference. Recently, pointing LSTM is designed to point the novel objects (Li et al., 2019) by balancing generation and copying words.

Diverse Image Description Generation is explored to incorporate diversity in the generated captions. Few approaches (Dai et al., 2017; Shetty et al., 2017) has leveraged adversarial training, while (Vijayakumar et al., 2016) used diverse beam search to decode diverse image captions in English. Approaches were also proposed to describe images from cross-domain (Chen et al., 2017).

Controllable Image Description Generation control and select the objects in an image to generate descriptions. Initially, (Yin & Ordonez, 2017) generated layout from images, while (Wang et al., 2018b) count image objects to produce multiple captions for a given image. Further, control signal is used to make the image captioning more controllable and to generate diverse captions. (Cornia et al., 2018) used either a sequence or a set of image regions. Also, chunks of the generated sentences are explicitly grounded on regions. Furthermore, instead of making captions only diverse, (Deshpande et al., 2018) make the generated descriptions accurate.

2.1.2 Video Description Generation

Going beyond images, the goal of video captioning is to comprehend the spatio-temporal information in a video for generating either single or multiple textual descriptions.

Global Video Description Generation approaches (Motwani & Mooney, 2012; Regneri et al., 2013) initially started by grounding sentences that describe actions in the visual information extracted from videos. It is further expanded into generating global natural language descriptions for videos with various approaches such as leveraging latent topics (Das et al., 2013), corpora knowledge (Krishnamoorthy et al., 2013), graphical models (Rohrbach et al., 2013) and sequence-to-sequence learning (Venugopalan et al., 2014, 2015; Donahue et al., 2015; Srivastava et al., 2015; Xu et al., 2016; Ramanishka et al., 2016; Jin et al., 2016). Figure 3 summarizes the description generation for the complete video. Aforementioned approaches leverage training datasets with limited objects, however recognition and description of entities and activities in real-world videos is hard. Generating natural language descriptions for such videos is also addressed with a factor graph by combining visual detection with language statistics (Thomason et al., 2014).

Further, sequence-to-sequence based approaches are improved with external corpora (Venugopalan et al., 2016) and also attention with various techniques such as soft-attention (Yao et al., 2015), multimodal fusion (Hori et al., 2017), temporal attention (Song et al., 2017), semantic consistency (Gao et al., 2017) and residual connections (Li et al., 2018). Apart

from attention based methods, novel architectures are also explored such as incorporation of semantic attributes learned from videos (Pan et al., 2017), ensemble based description generator networks (Shetty et al., 2018) and encoder-decoder-reconstructor which leverages both the forward and backward flows i.e., (video to description) and (description to video) respectively for the video captioning (Wang et al., 2018). Multi-faceted attention is also explored to select the most salient visual features or semantic attributes, and then overall sentence is generated (Long et al., 2018).

Apart from architecture improvements, different machine learning approaches are explored as well, video captioning is seen from a multi-task learning scenario by sharing knowledge between two related tasks such as temporal- and context-aware video combined with entailment generation task (Pasunuru & Bansal, 2017a). It is further approached from the reinforcement learning by providing entailment rewards (Pasunuru & Bansal, 2017b) and to address the description generation for multiple fine-grained actions (Wang et al., 2018b).

In the following, we explore some of the related ideas which expand the scope of video description generation.

Dense Video Description Generation work in the similar manner to dense image description generation, dense video description generation aims to address the fine-grained video understanding by addressing two sub-problems i.e., localizing events in an video and then generating captions for these localized events (Zhou et al., 2018c; Xu et al., 2019). Further, (Zhou et al., 2018a) extended earlier research by explicitly linking the sentence to corresponding bounding box in one of the frames of a video by annotating each noun phrase observed in the sentence. Incorporation of the background knowledge for video description generation is also explored as another line of research (Whitehead et al., 2018). However, the core challenge which is still remaining is automatic evaluation of the video captioning and is currently studied from the perspective of direct assessment (Graham et al., 2018).

Movie Description Generation saw video description generation from an different perspective where movie clips are used as input. Initially, aligning books to movies (Tapaswi et al., 2015; Zhu et al., 2015) is used to generate story like explanations. Furthermore, movie descriptions (Rohrbach et al., 2015) were directly created by transcribing audio description concentrating on precisely describing what is shown in the movie.

2.2 Visual Storytelling

A natural way to extend the visual description generation where a single image is described with a sentence to something like a paragraph. However, considering sequence of images or frames (in video) for generating a paragraph which is story like is much more viable and attractive compared to generation of paragraph from a single image.

2.2.1 Image Storytelling

The aim of image storytelling is to generate stories from sequences of images. Although, sequence of images can been perceived in the form of a video. In contrast to videos, consecutive images in the streams can have sharp changes of visual content, which cause the abrupt discontinuity between consecutive sentences (Park & Kim, 2015). Hence, it is seen as a sequential vision-to-language task (Huang et al., 2016) where images are not considered

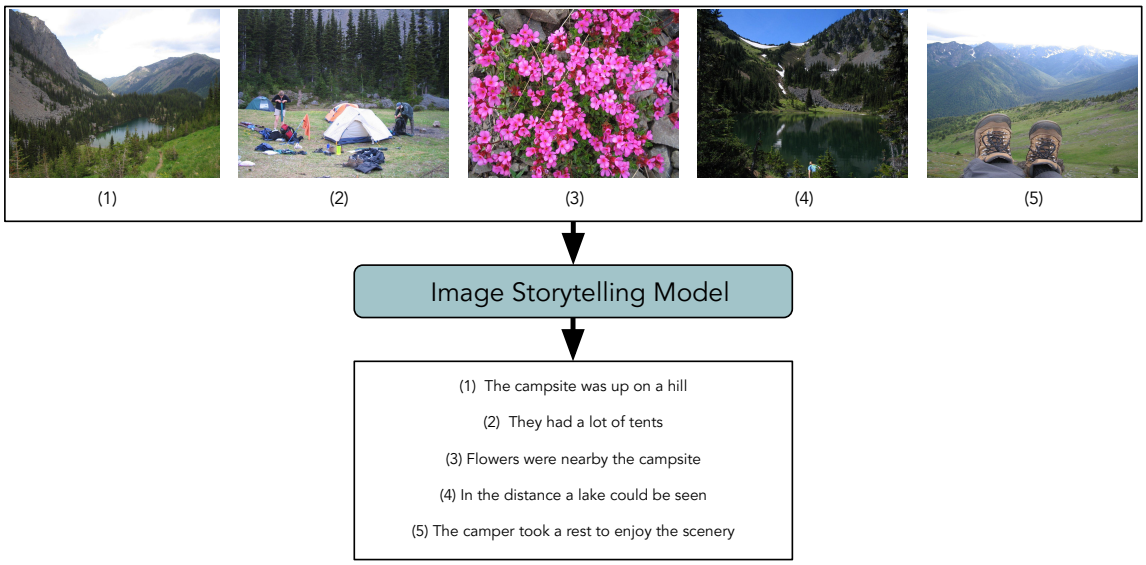

Figure 4: Given sequence of images, the image storytelling model generates a textual story in sequence.

in isolation. Figure 4 summarizes the image storytelling where a story in a sequence is generated.

Initially, semantic coherence in a photo stream is captured by reducing the visual variance. Further, semantic space is acquired by jointly embedding each photo with its corresponding contextual sentence such that their correlations are discovered (Liu et al., 2017). It is then improved by exploiting hierarchical architecture (Yu et al., 2017) and then further optimized by incorporating reinforcement learning with rewards (Wang et al., 2018a) for generating relevant and expressive narrative paragraphs. Instead of flat deep reinforcement learning, the hierarchically structured reinforced training is also studied (Huang et al., 2018) and has shown to achieve significantly better performance than the former. Similarly, (Wang et al., 2018a) used adversarial reward learning to learn an implicit reward function from human demonstrations to optimize policy search with the learned reward function.

Nevertheless, the standard way of narration suffers from repetitiveness where the same objects or events will undermine a good story structure. Hence, inter-sentence diversity was explored with diverse beam search to generate more expressive stories (Hsu et al., 2018). The task is also observed from a different perspective where given a jumbled set of aligned image-description pairs that belong to a story, the task is to sort them such that the output sequence forms a coherent story (Agrawal et al., 2016).

While earlier research addresses only natural images, (Li et al., 2019) incorporated medical domain knowledge to generate realistic and accurate descriptions for medical images.

2.2.2 Video Storytelling

In comparison with image storytelling which only deal with sequence of images, the aim of video storytelling is to generate coherent and succinct stories for long videos. However, video storytelling is less explored in contrast with image storytelling. (Li et al., 2018)

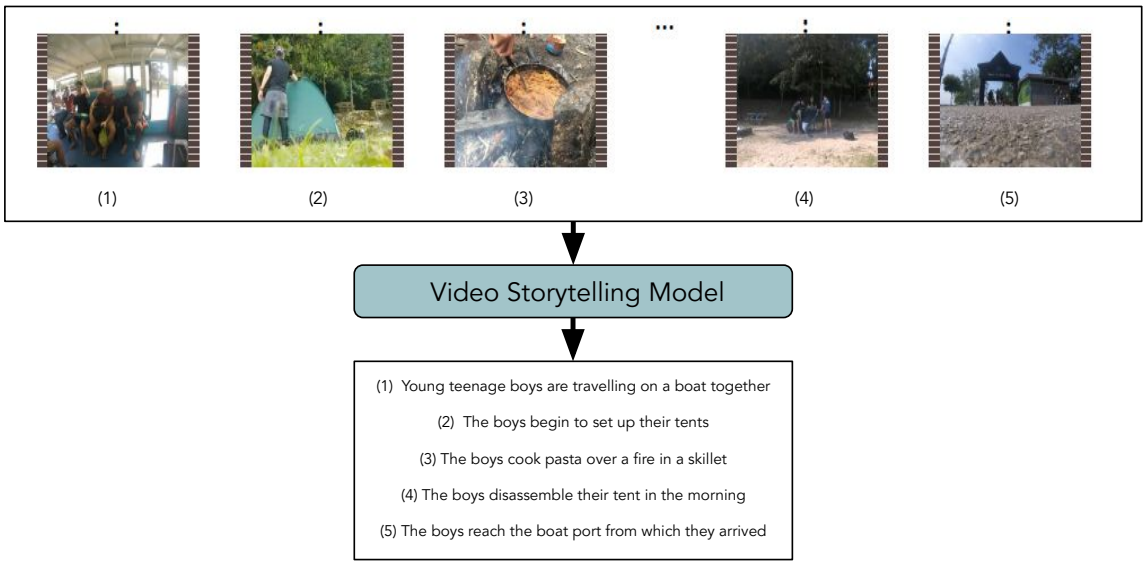

Figure 5: Given Video Frames (adopted from (Li et al., 2018)), the video storytelling model generates a textual story in sequence.

first introduced the problem of video storytelling to address challenges such as diversity in the story and inherent complexity of the video. They introduce residual Bidirectional RNN (BiRNN) for leveraging context and a narrator model with reinforcement learning. Further, (Gella et al., 2018) created multi-sentence video description dataset (VideoStory) resembling stories from social media videos. It was aimed to support disability or network bandwidth challenges. Figure 5 summarizes the video storytelling where a story in a sequence is generated when given an input video.

Nevertheless, this task can be seen very close to the well researched area of video summarization (Ma et al., 2002) using only videos.

2.3 Visual Question Answering

The goal of visual question answering (VQA) is to learn a model which comprehends the visual content at both global- and local-level for finding an association with pairs of questions and answers in the natural language form. Both images and videos are used as visual information for VQA.

2.3.1 Image Question Answering

The aim of image question answering (image Q&A) is to answer text-based questions about images. Earlier research has designed different algorithms and datasets to address the challenge. Initially, works (Malinowski & Fritz, 2014; Malinowski et al., 2015; Geman et al., 2015) considered image Q&A as the Visual Turing Test, where the expectation is to incorporate the human-level abilities for semantically accessing the visual information to answer different questions. It is then further improved as a fill-in-the-blank task (Yu et al., 2015), where the goal of the system is focused on multiple-choice question-answering for images.

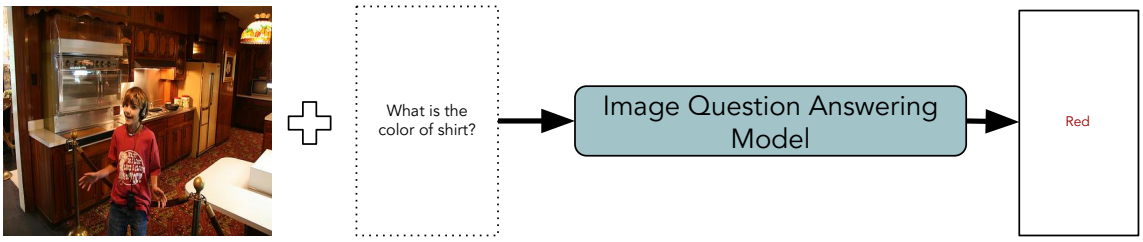

Figure 6: Given an Image and Question, the Image Question Answering Model produce an answer to it.

Also, it is expanded to address both multilingual (Gao et al., 2015) and automatic question generation such that description of sentences are converted into questions (Ren et al., 2015a). However, it lacked the natural-language questioning ability of humans. Hence, a broader task was proposed with an aim of addressing the open-ended image Q&A (Antol et al., 2015; Agrawal et al., 2017), where the challenge was to ask a free-form natural-language question about an image to make the system answer the question. Figure 6 presents the task where a free-form question about an image is asked to attain an answer.

However, designing such a system can contain several other challenges such as coming up with strong baselines (Jabri et al., 2016). To address it, binary image Q&A (Zhang et al., 2016) was explored by providing complementary images for the abstract scenes. Task is observed from the perspective of visual verification of concepts inquired in the questions. Some of the questions where understood as a loose, global association between Q&A sentences and images. Hence, more confined and dedicated tasks was created for relating local regions in the images (Zhu et al., 2016) by addressing object-level grounding. Some approaches (Zhang et al., 2018a) concentrated only on counting objects in natural images. There are many methods that are proposed to address the challenging image Q&A task. The details about different methods are covered in earlier surveys (Kafle & Kanan, 2017; Wu et al., 2017b). However, we briefly present new methods that are published after those surveys.

Recent works aim at interpretability/explainability by overcoming priors (Agrawal et al., 2018), concentrate better on the image to extract relevant information (Goyal et al., 2019), human-interpretable rules that gives better insights (Manjunatha et al., 2018), and cycle-consistency (Shah et al., 2019a). While, other works target at understanding the text inside an image to answer and reason about it (Singh et al., 2019). Other works aimed to include outside knowledge (Marino et al., 2019) into image Q&A to support real-world knowledge aware question answering (Shah et al., 2019b).

2.3.2 Video Question Answering

The goal of the video question answering (video Q&A) is to answer text-based questions about videos and is less explored than the image Q&A. Nevertheless, there are few works which explored this spatio-temporal domain. Initially, joint parsing of the videos and corresponding text to answer queries is explored (Tu et al., 2014). Further, an open-ended Movie Q&A (Tapaswi et al., 2016) having multiple-choice question pairs is designed to solve challenging questions that require semantic reasoning over a long temporal domain.

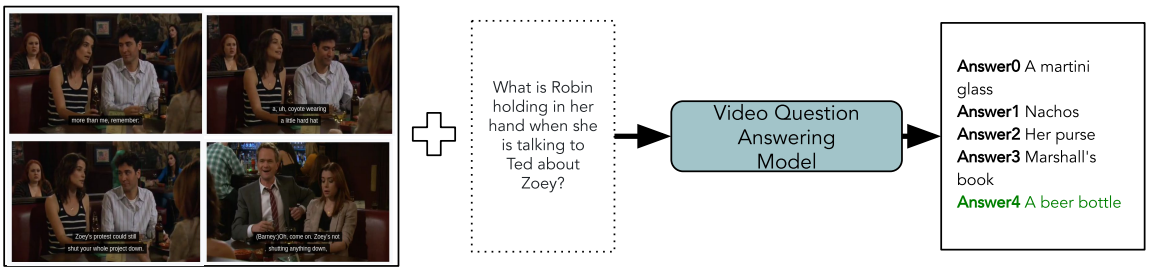

Figure 7: Given a Video (represented as sequence of frames from TV Q&A dataset) and question, the Video Question Answering model find the right answer from Multiple Options.

Further, to limit the involvement of crowd workers, the task was modified using fill-in-the-blank questions (Zhu et al., 2017; Mazaheri et al., 2017) and is automatically generated from the different manually created video description datasets. Other works (Zeng et al., 2017) have modified it to support answering free-form natural language questions. Furthermore, open-ended video question answering is also addressed with methods such as spatio-temporal attentional encoder-decoder learning framework (Zhao et al., 2017). Interest has been shown to jointly address multiple tasks that handle video and language. High-level concept words (Yu et al., 2017b) are detected such that they can be integrated with any video and language models addressing fill-in-the blank and multiple-choice test. Spatio-temporal reasoning from videos to answer questions is also addressed by designing a both spatial and temporal attention (Jang et al., 2017).

Recently, due to large interest in video Q&A, similar to Movie Q&A, six popular TV shows were used to create a dataset, where questions are compositional (Lei et al., 2018). The TV Q&A dataset made the proposed multi-stream models to jointly localize relevant moments within a clip, comprehend subtitle-based dialogue, and recognize relevant visual concepts. Furthermore, spatio-temporal grounding (Lei et al., 2019) is done to link depicted objects to visual concepts in questions and answers. Figure 7 present the task where question about the video is asked to detect answer from multiple choices.

2.4 Visual Dialog

2.4.1 Image Dialog

The goal of the Image dialog task is to create an AI agent for holding a dialog with humans in natural language about the visual content (Das et al., 2017a) represented by an image. To be specific, given an image, a history about dialogues, and a question about the image. The goal of the agent is to ground the question in image, infer context from the history, and answer the question accurately. However, the problem can be observed from different perspective where the goal of the system is to locate an unknown object in an image by asking a sequence of questions (de Vries et al., 2017) or natural-sounding conversations about a shared image (Mostafazadeh et al., 2017). Figure 8 summarizes the task.

Further, a standard agent is extended to have a question and answer bot to cooperate between each other for guessing images (Das et al., 2017b). To counter generic responses in dialog generation, knowledge transfer from Dialog generation is explored with the dis-

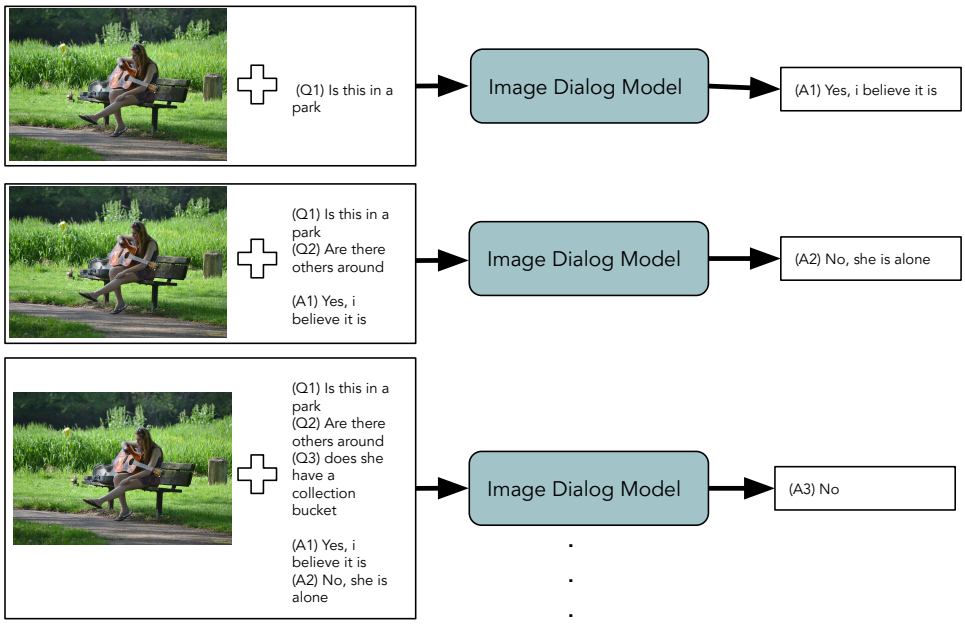

Figure 8: Given an Image, Question and the Dialog history, the Image dialog model generates answer based on it.

criminative Dialog module trained to rank a list of candidate human responses (Lu et al., 2017). However, other approaches constrained themselves to a domain proposing end-to-end optimization (Strub et al., 2017). (Seo et al., 2017) proposed attentive memory that exploits visual attention in the past to resolve the current reference. Recently, reinforcement learning and Generative Adversarial Networks (GANs) were also explored to generate more human-like responses to questions in the image based Dialog (Wu et al., 2018). Dialog is also seen from the perspective of a system which asks questions, and demonstrate how visual dialog can be generated from discriminative question generation and answering (Jain et al., 2018). Furthermore, co-reference resolution is also explored (Kottur et al., 2018) to bridge the gap between nouns and pronouns by usage of modules that form explicit, grounded, co-reference resolution at word-level.

Recently, a novel attention mechanism called recursive visual attention (Niu et al., 2019) was proposed to resolve visual co-reference for visual dialog by browsing the dialog history. Another approach (Zheng et al., 2019) formalized the task as inference in a graphical model with partially observed nodes and unknown graph structures i.e., relations in dialog. Further, (Guo et al., 2019) extended one-stage solution with a two-stage solution by building a image-question-answer synergistic network to value the role of the answer for precise visual dialog.

2.4.2 Video Dialog

The aim of video dialog is to leverage scene information containing both audio (can be transcribed as subtitles) and visual frames to hold a dialog with humans in natural language about the content (Alamri et al., 2019b, 2019a). A successful system is expected to ground

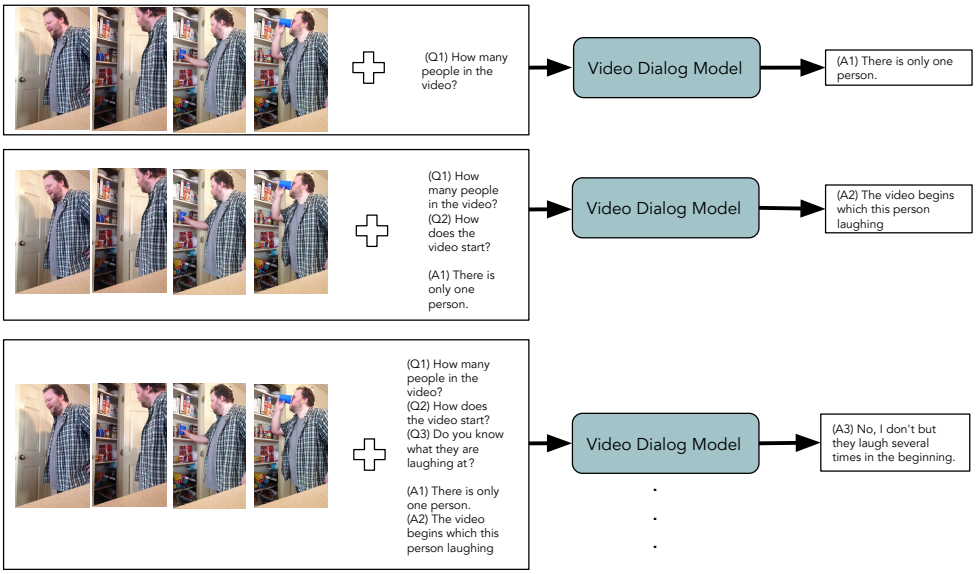

Figure 9: Given a video (represented as sequence of frames), Question and Dialog history, the Video dialog model generates answers based on it.

concepts from the question in the video while leveraging contextual cues from the dialog history. Figure 9 summarizes the task.

Several approaches were proposed to address it, initially multimodal attention-based video description features are used to improve dialog (Hori et al., 2019). Further, a novel baseline (Schwartz et al., 2019) analyzes components such as data representation extraction, attention and answer generation show that there can relative improvements as proposed to other approaches.

2.5 Visual Reasoning

2.5.1 Image Reasoning

The goal of image reasoning is to answer sophisticated queries by reasoning about the visual world. Initial works (Johnson et al., 2017a) aimed at designing diagnostic tests going beyond benchmarks such as VQA. They reduced the biases by having detailed annotations describing the kind of reasoning each question requires. It is also observed in VQA models that they struggle on tasks such as comparing the attributes of objects, or compositional reasoning where novel attribute combinations needs to be recognized. Further, a novel approach (Johnson et al., 2017b) is proposed by using a program generator to construct an explicit representation of the reasoning process, and an execution engine that executes the resulting program to produce an answer. Then, end-to-end module networks (Hu et al., 2017) are proposed which learn to reason by directly predicting instance-specific network layouts without the aid of a parser as used in neural module networks. (Santoro et al., 2017) go beyond and propose Relation Networks (RNs) as a simple plug-and-play module to solve the problem of visual reasoning. RNs are further used to learn relation-aware visual features for content based image retrieval (Messina et al., 2018) and also Multi-Relational

Network (Chang et al., 2018). Furthermore, global context reasoning (Cao et al., 2018) is explored for better aligning image and language domains in diverse and unrestricted cases.

Recent approach (Perez et al., 2018) introduces a general-purpose conditioning method Feature-wise Linear Modulation (FiLM) layers which influence neural network computation via a simple, feature-wise affine transformation based on conditioning information. FiLM is further modified (Strub et al., 2018) to generate parameters of FiLM layers going up the hierarchy of a convolutional network in a multi-hop fashion rather than all at once. Cascaded Mutual Modulation (CMM) (Yao et al., 2018) proposed with a end-to-end visual reasoning model also uses the FiLM technique to enable the textual/visual pipeline to mutually control each other. Another approach modifies Neural modular networks (Hu et al., 2018) such that it performs compositional reasoning by automatically inducing a desired sub-task decomposition without relying on strong supervision. (Mascharka et al., 2018) proposed a set of visual-reasoning primitives which, when composed, manifest as a model capable of performing complex reasoning tasks in an explicitly interpretable manner. Also, in the context of interpretable learning framework learning-by-asking (LBA) (Misra et al., 2018b) is proposed to closely mimic natural learning with the goal to make it more data efficient than the traditional VQA setting. Further, compositional attention networks (Hudson & Manning, 2018) are proposed which is a fully differentiable neural network architecture, designed to facilitate explicit and expressive reasoning. The goal of the architecture is towards a design that provides a strong prior for iterative reasoning, allowing it to support structured learning, as well as generalization from a modest amount of data.

Recently, neural-symbolic visual question answering (Yi et al., 2018) is proposed to combine deep representation learning with symbolic program execution. It first recovers structural scene representation from the image and a program trace from the question. An extension of it achieved with a Neuro-Symbolic Concept Learner (NS-CL) (Mao et al., 2019) that learn visual concepts, words, and semantic parsing of sentences without explicit supervision. It just learns by simply looking at images and reading paired questions and answers. Further, a multimodal relational network (MuRel) (Cadene et al., 2019) is proposed to learn end-to-end reasoning over real images. Additionally, (Aditya et al., 2019) used spatial knowledge to aid in visual reasoning. They proposed a framework that combines knowledge distillation, relational reasoning and probabilistic logical languages. Existing diagnostic tests are further modified with referring expressions to handle bias (Liu et al., 2019) and with structural, relational, and analogical reasoning in a hierarchical representation (Zhang et al., 2019). Explainable and explicit neural modules (Shi et al., 2019) are also explored with scene graphs. Objects as nodes and pairwise relationships as edges are used for explainable and explicit reasoning with structured knowledge.

Further expanding the scope, (Andreas et al., 2016a, 2016b) exploit the compositional linguistic structure of complex questions by forming neural module networks which query about the abstract shapes observed in an image. Improvement is further seen in how images are interpreted i.e., compositional question answering (Hudson & Manning, 2019) is addressed with scene graph structures on real-world images going beyond abstract shapes. Figure 10 present the task of reasoning about real world images.

Reasoning is also extended to cognition for understanding the commonsense observed from images with commonsense reasoning (Zellers et al., 2019).

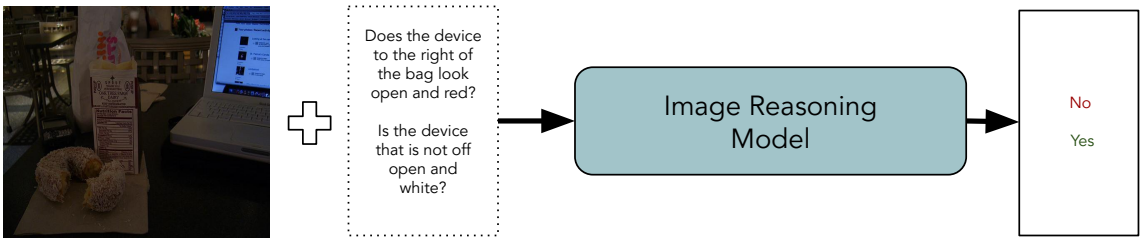

Figure 10: Given a real-world Image and Question, the Image Reasoning model reasons about the question to produce an answer.

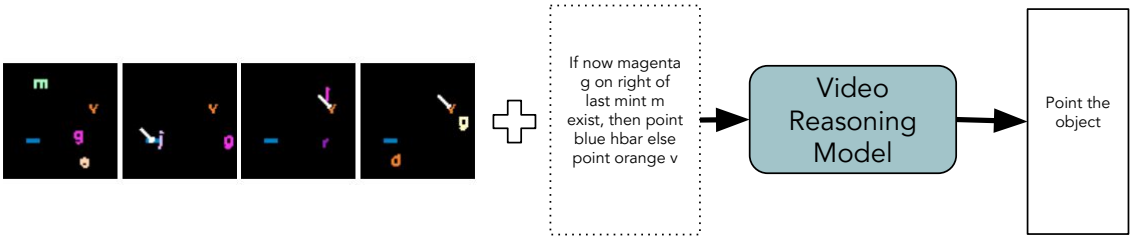

Figure 11: Given a Video (represented as sequence of synthetic 2D scenes from (Yang et al., 2018)) and Question, the Video Reasoning model reasons about to perform the task presented to it in the question.

2.5.2 Video Reasoning

Compared to image reasoning, video reasoning is in its nascent stages. Still there is no clear goal defined. However, for video reasoning, a configurable visual question and answer (COG) (Yang et al., 2018) is designed to parallel experiments in humans and animals. The goal of COG is to address problems relating to visual and logical reasoning and memory. To be specific, the task is aimed to deduct the correct answer while taking into account changes of the scene i.e., from both spatial and temporal perspective. Figure 11 present the task of temporal reasoning about synthetic 2D scenes resembling video input.

Further, (Haurilet et al., 2019) addressed both image and video reasoning by introducing the concept of a question-based visual guide to constrain the potential solution space by learning an optimal traversal scheme. In their approach, the final destination nodes alone are used to produce the answers.

2.6 Visual Referring Expression

2.6.1 Image Referring Expression

In a natural environment, people use referring expressions to unambiguously identify or indicate particular objects. This is usually done with a simple phrase or using a larger context e.g. sentence. Having a larger context provides better scope for avoiding ambiguity. It is important because the referential expression can easily map the target object. However, there can also be other possibility where people are asked to describe a target object based on its surrounding objects.

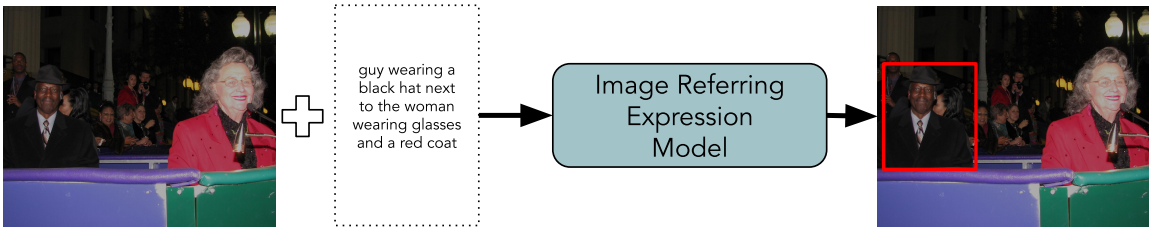

Figure 12: Given an Image and Referring Expression, the Image Referring expression comprehension identify it in the image using bounding boxes.

This provide us with two different possibilities for the visual referring expression task. In the first scenario, referring expression deals with generation where an algorithm generates a referring expression for a given target object present in a visual scene. While in the second scenario, the referring expression used to perform comprehension where an algorithm localizes the object in an image described by a given referring expression. Given these tasks, different approaches are proposed for referring expression generation (Golland et al., 2010; Mitchell et al., 2013), comprehension (Kazemzadeh et al., 2014) and both (Mao et al., 2016; Yu et al., 2016). Note that there is difference from referring expression tasks in contrast with grounding of free-form textual phrases (Rohrbach et al., 2016b) in an image. Figure 12 shows an example of the referring expression comprehension.

Referring Expression Generation approaches (FitzGerald et al., 2013) proposed initially observed the problem from a density estimation perspective where the goal is to learn distributions over logical expressions identifying sets of objects in the world. Further, a comprehension guided referring expression generator is designed (Luo & Shakhnarovich, 2017) by using the comprehension module trained on human-generated expressions for generating referring expressions.

Referring Expression Comprehension was explored (Nagaraja et al., 2016) to integrate context between objects. Later, techniques such as multiple instance learning (MIL) were used to explore context regions and max-margin based MIL objective functions for training. Further, (Hu et al., 2016) leveraged a natural language query of the object to localize a target object using a spatial context recurrent convnet (SCRC) model. It works as a scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information. This explicit modeling of the referent and context region pairs has proven useful. Approaches such as compositional modular networks (Hu et al., 2017b) analyzed referential expressions by identifying entities and relationships mentioned in the input expression and grounding them all in the scene. This showed to effectively inspect local regions and pairwise interactions between them. A modular approach is also explored where three modular components related to subject appearance, location, and relationship to other objects is used to model with Modular Attention Network (Yu et al., 2018). It has shown to focus well on the subjects and their relationships. Approaches such has GroundNet (Cirik et al., 2018) have leveraged syntactic analysis of the input referring expression to build a dynamic computation graph of neural modules that defines architecture for performing localization. Variational models are also studied for

referential expression comprehension where variational Bayesian methods called variational context (Zhang et al., 2018b) is used to solve the problem of complex context modeling. It has shown to exploit the relation between the referent and context thereby reducing the search space of context. Furthermore, an accumulated attention mechanism (Deng et al., 2018) is proposed to accumulate the attention for useful information in image, query, and objects. It has shown to reduce the redundancy and noise issues in other approaches.

Recently, a cross-modal relationship extractor (CMRE) and a gated graph convolutional network (GGCN) combined into cross-modal relationship inference network (Yang et al., 2019). CMRE has shown to highlight objects and relationships which have connections with a given referring expression. While, GGCN computes multimodal semantic contexts by fusing information from different modes and propagates multimodal information in the structured relation graph. Thinking from the natural language understanding perspective, a Recursive Grounding Tree (Hong et al., 2019) is built to automatically compose a binary tree structure by parsing the referring expression to perform visual reasoning along the tree in a bottom-up fashion. It has shown to allow gradients from continuous score functions with a discrete tree construction. There are also interests in combining visual reasoning with referential expressions by creation of new dataset (Liu et al., 2019). Most of the above approaches use bounding box localization, additionally object segmentation (Liu et al., 2017) is also explored for referring expression comprehension.

Referring Expression Generation and Comprehension Few approaches have performed both generation and comprehension. Visual context (Mao et al., 2016; Yu et al., 2016) was initially used in referring expression models to find visual comparison to other objects within an image. It has shown significant improvements. Further, an unified framework (Yu et al., 2017a) is designed using speaker, listener, and reinforcer where the speaker generates referring expressions, the listener comprehends referring expressions, and the reinforcer introduces a reward function to guide sampling of more discriminative expressions. It has shown to benefit the tasks from the discriminative reinforcer's feedback. The role of attributes (Liu et al., 2017) is also studied to show that is helps in ambiguation when referring to a particular object.

2.6.2 Video Referring Expression

When comparing to image referring expression, video referring expression is less explored. But, there is surge in interest to deal with spatial-temporal contexts and motion features observed in videos. However, most of the work is concentrated only on one variant of image referring expression i.e., comprehension. (Balajee Vasudevan et al., 2018) used stereo videos to exploit richer, more realistic temporal-spatial contextual information along with gaze cues for referring expression comprehension. Figure 13 shows an example of the video referring expression comprehension. Another approach by (Khoreva et al., 2018) explored Language Referring Expressions to point to the object in a video to achieve object segmentation.

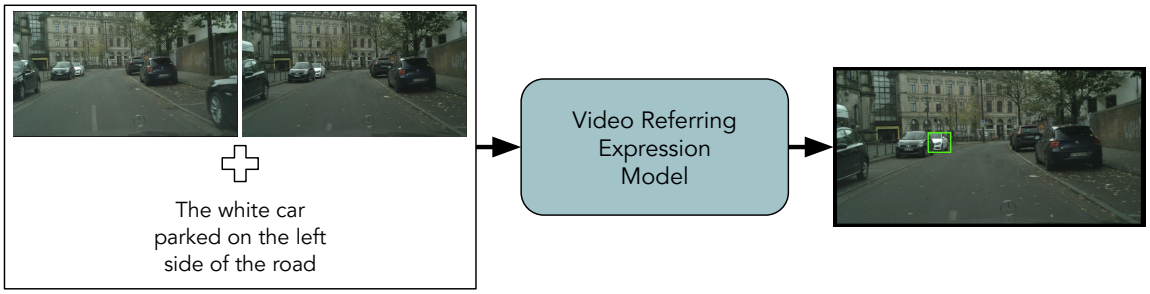

Figure 13: Given a Video (represented as sequence of frames from (Balajee Vasudevan et al., 2018)) and the Referring Expression, the Referring Expression Comprehension model identifies it in the video using bounding boxes.

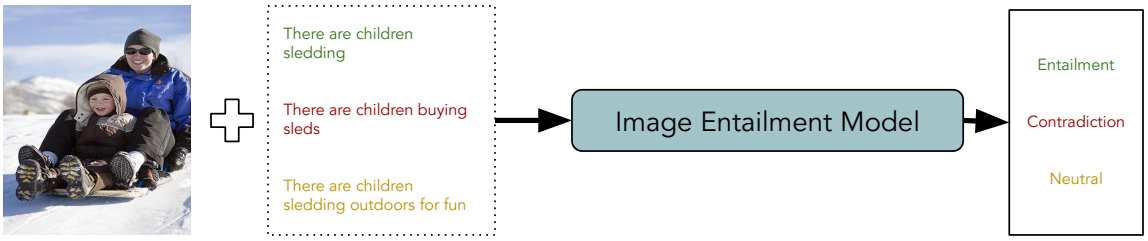

Figure 14: Given an image *premise* and a natural language text as *hypothesis*, the Image Entailment model predicts whether the hypothesis is an entailment, contradiction or neutral by understanding evidence observed in the image.

2.7 Visual Entailment

2.7.1 Image Entailment

Addressing the drawbacks of VQA and the visual reasoning which deal with similar objects and sentence structures. (Vu et al., 2018) initially proposed a visually-grounded version of the Textual Entailment task where image is augmented to textual premise and hypothesis. However, it is further refined by (Xie et al., 2019) to predict whether the image semantically entails the text, given image-sentence pairs where premise is defined by an image instead of a natural language sentence. Figure 14 summarizes the task, where image premise and textual hypothesis are used by the Image Entailment model to predict whether the hypothesis is an entailment, contradiction or neutral.

2.8 Language-to-Vision Generation

2.8.1 Language-to-Image Generation

Sentence-level Language-to-Image Generation goal is to generate images conditioned on the natural language descriptions. It is considered as a fundamental problem in many applications. The success of generative adversarial networks (GAN) (Goodfellow et al., 2014) has shown to generate interesting images of specific categories, such as room interiors, album covers, and faces (Radford et al., 2015). This has led the interest in bridging the gap between the natural language text and image modeling. Figure 15 shows that the

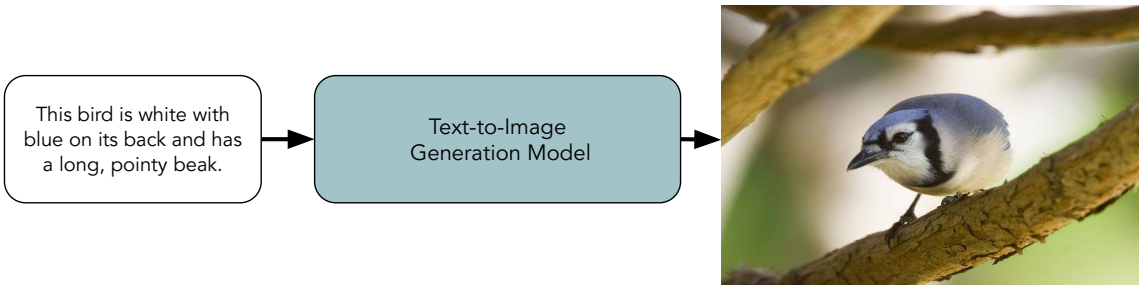

Figure 15: Given a natural language description, the Language-to-Image model generates an image conditioned on the description.

natural language description is used to generate an image with a text-to-image generation model.

Initially, alignDRAW (Mansimov et al., 2015) was introduced to iteratively draw patches on a canvas, while attending to the relevant words in the description. Further, visual concepts are translated from characters to pixels (Reed et al., 2016a) with a conditional GAN. It was further improved (Reed et al., 2016b) by the taking instructions about what content to be drawn in which location achieving high-quality image generation. Models which were developed to condition on classes for image generation (Nguyen et al., 2017) are also explored to generate images. However, their quality of image generation is much lower than when conditioning on classes. Most of the earlier approaches generated low-resolution images. Very close to this approach is text conditioned auxiliary classifier GAN (TAC-GAN) (Dash et al., 2017) which conditions images on both sentence and class information which has shown to improve their structural coherence. To generate images with high resolution, several GANs are stacked together as stackGAN (Zhang et al., 2017, 2018a) using the global sentence representation. This helped to generate images of different sizes. To over come the bottleneck of global-level sentence representation, attention based GAN as AttGAN (Xu et al., 2018) is introduced to capture the fine-grained details at different sub-regions of the image. It pays attention to the relevant words in the natural language description.

Further, a hierarchical approach (Hong et al., 2018) is explored by inferring semantic layout of the image. Instead of learning a direct description to an image mapping, the generation process is decomposed into multiple steps. First a semantic layout from the text is constructed by the layout generator and then the layout is converted to an image by the image generator. Other kind of approaches such as HDGAN (Zhang et al., 2018b) aim to accompany the hierarchical adversarial objectives inside the network to regularize mid-level representations and assist generator training to capture the complex image information. This has shown to generate images with high resolutions.

Later, instead of dealing with natural-language descriptions (Johnson et al., 2018) used image specific scene graphs enabling explicitly reasoning about objects and their relationships. Further, for getting better high resolution images, coarse-resolution features are taken as input and perceptual pyramid adversarial network (PPAN) is introduced to directly synthesize multi-scale images conditioned on texts in an adversarial way (Gao et al., 2019). Another approach named MirrorGAN (Qiao et al., 2019) targets the main goal of visual realism and semantic consistency for generating image from the text. It proposes global-local

attentive and semantics preserving framework where the image generated from the text is further used to generate the text back. This has shown to semantically align with the given text and generated description.

In the following, we explore some of the related ideas which expand the scope of language-to-image generation.

Image Manipulation take a different path from the earlier benchmark approaches about image generation, the TAGAN (Nam et al., 2018) was introduced to generate semantically manipulated images while preserving text-irrelevant contents. Here, the generator learns to generate images where only regions that correspond to the given text are modified. Another interesting approach is to have an interactive system that generates an image iteratively. There are also other variations where the source image is manipulated via natural language dialogue (Cheng et al., 2018).

Fine-grain Image Generation uses a recurrent image generation model (El-Nouby et al., 2018) to take into account both the generated output up to the current step as well as all past instructions for generation. This has shown to add new objects, apply simple transformations to existing objects, and correct previous mistakes. Earlier research never concentrated on fine-grained generation of images i.e., localizing objects. Recently, control of the location of individual objects within an image is achieved (Hinz et al., 2019) by adding an pathway for iteratively applying them at different locations specified by the bounding boxes to both generator and the discriminator.

Sequential Image Generation approach StoryGAN (Li et al., 2018a), based on the sequential conditional GAN concentrate on story by generating a sequence of images, when given a multi-sentence paragraph. Termed as story visualization, it behaves exactly opposite to image storytelling and have shown to generate images with high quality and also achieving contextual consistency.

2.8.2 Language-to-Video Generation

The goal of language-to-video generation is to mimic language-to-image generation by considering the temporal aspect. However, language-to-video generation requires a stronger conditional generator than what is generally required for the language-to-image generation. This is because of the increase in dimensionality. To address this challenge, a conditional generative model is trained (Li et al., 2018b) to extract both static and dynamic information from text which combines variational autoencoders (VAE) (Kingma & Welling, 2013) with GAN. Figure 16 shows that the natural language description is used to generate a video with text-to-video generation model.

Another novel approach is to generate video from script. The composition, retrieval and fusion network (Craft) model (Gupta et al., 2018) is proposed which is capable of learning knowledge from the video-description data and applying it in generating videos from novel captions. It has shown that the Craft model performs better than the direct pixel generation approaches and generalizes well to unseen captions and to video databases with no text annotations.

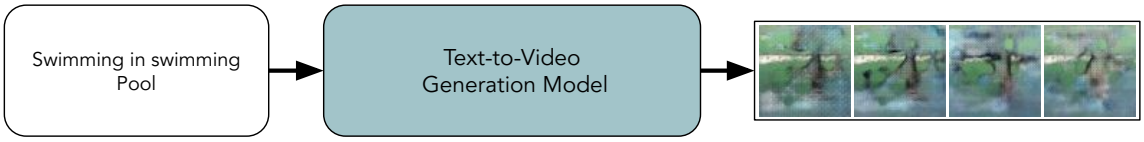

Figure 16: Given a natural language description, the Language-to-Video model generates a video (represented as sequence of frames from (Li et al., 2018b)) conditioned on the description.

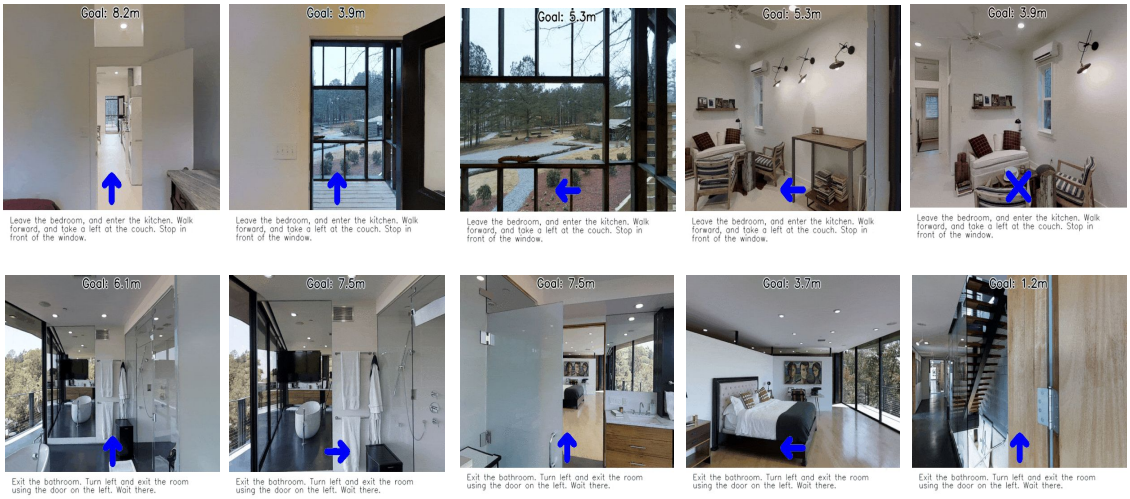

Figure 17: Given an Image and Language instructions (represented with sequence of images from (Anderson et al., 2018)), the Image-and-Language Navigation model is used to carryout the navigation of an agent in an environment (indicated by arrows).

2.9 Vision-and-Language Navigation

2.9.1 Image-and-Language Navigation

Most of the works proposed for vision-and-language navigation (VLN) are proposed for photo-realistic images forming 3D environments. The goal of the image-and-language navigation (ILN) is to enable an agent or a robot to carry out navigation in an environment given by the photo-realistic image views by interpreting natural language instruction (Anderson et al., 2018). This requires the agent/robot to simultaneously process both vision and language and navigate from a source to a target location. Figure 17 summarizes the task.

Initially, sequence-to-sequence model was proposed to address such a challenge where the student forcing approach achieved promising results in previously explored environments. The approach (Wang et al., 2018) which was proposed later has integrated a module to combine model-based and model-free reinforcement learning techniques to better generalize to unseen environments. Furthermore, the reinforced cross-modal matching approach (Wang et al., 2018c) is proposed which enforces cross-modal grounding both locally and globally via reinforcement learning.

ILN is also seen from the perspective of search on a navigation graph (Ma et al., 2019b) and a progress monitor is developed as a learnable heuristic for search. It is also improved by leveraging a visual-textual co-grounding attention mechanism to better align the instruction and visual scenes, and incorporates a progress monitor to estimate the agent's current progress towards the goal (Ma et al., 2019a). Another substantial improvement is seen in training and action space with a embedded speaker model (Fried et al., 2018). New instructions are synthesized for data augmentation and to implement pragmatic reasoning for evaluating how well candidate action sequences explain an instruction. Improving over earlier approaches that make local action decisions or score entire trajectories using beam search, the novel approach of FAST framework (Ke et al., 2019) balances local and global signals when exploring the environment allowing it to act greedily, but use global signals to backtrack when necessary. Also, (Tan et al., 2019) explore a generalizable navigational agent by training it in two stages. Where in the first stage, mixed imitation and reinforcement learning is combined, while in the second stage, fine-tuning is performed via newly-introduced "unseen" triplets.

ILN is also seen from the perspective of the visual question answering (see Section 2.3) that requires navigation to answer questions. Embodied Question Answering (Das et al., 2018a, 2018b) is explored with an agent that is spawned at a random location in a 3D environment and asked a question. For answering the question, the agent navigates through the 3D environment finding the information observed in the question. Also, it is explored from interactive question answering (Gordon et al., 2018) and grounded dialog (de Vries et al., 2018). Other set of approaches (Misra et al., 2018a) aims to map instructions to actions in 3D Environments with visual goal prediction.

2.10 Multimodal Machine Translation

2.10.1 Machine Translation with Image

The aim of MMT (Specia et al., 2016; Hitschler et al., 2016; Elliott et al., 2017; Barrault et al., 2018) is to translate source language sentences that describe an image into target language. However, for any given image one or more descriptions in the source language is plausible, thus having a possibility to propose different variants of the task. First, a single source translation task is proposed, in which a source language image description is translated to the target language with additional cues from the corresponding image. Figure 20 summarizes this approach, where an image is accompanied with a English language description to be translated into German description.

Second, a target language description generation task with additional source language cues i.e., multiple source language descriptions of the same image seen as multisource MMT. Figure 19 summarizes the approach, where an image is accompanied with English, French and Czech descriptions used to generate the German translation.

Different approaches are proposed to handle single source MMT by associating visual and textual features with multimodal attention (Huang et al., 2016). Further, novel approaches are proposed (Calixto et al., 2017) in which a doubly-attentive decoder incorporates visual features and bridge the gap between image description and translation and global visual features for attention-based neural machine translation (Calixto & Liu, 2017). This is achieved

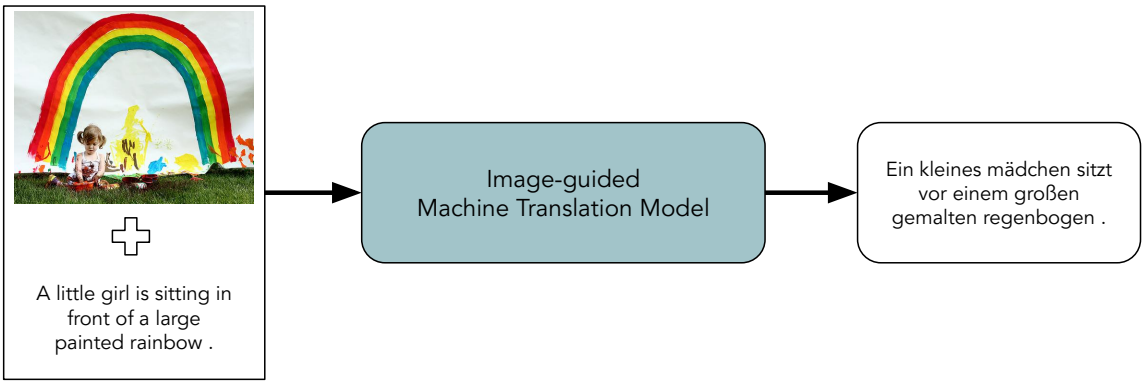

Figure 18: Given an Image and a single source language description, the Image-guided Machine Translation produce the target language description.

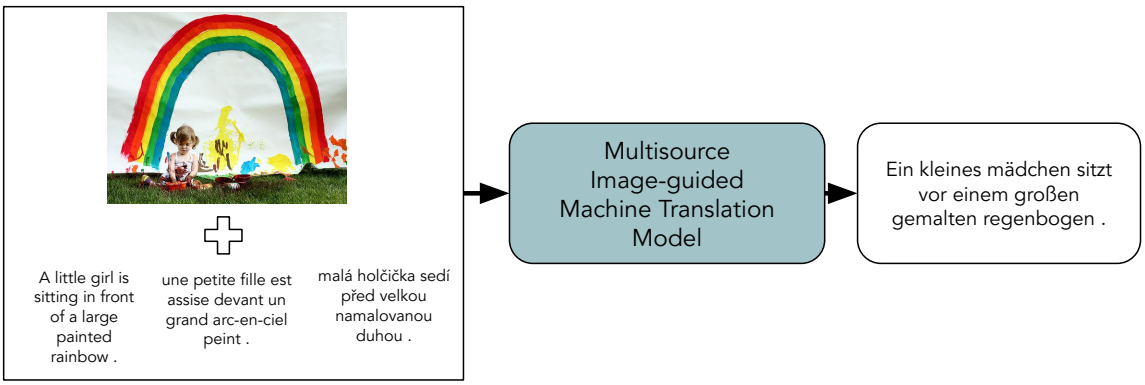

Figure 19: Given an Image and multiple source language descriptions, the Multisource Image-guided Machine Translation produce the target language description.

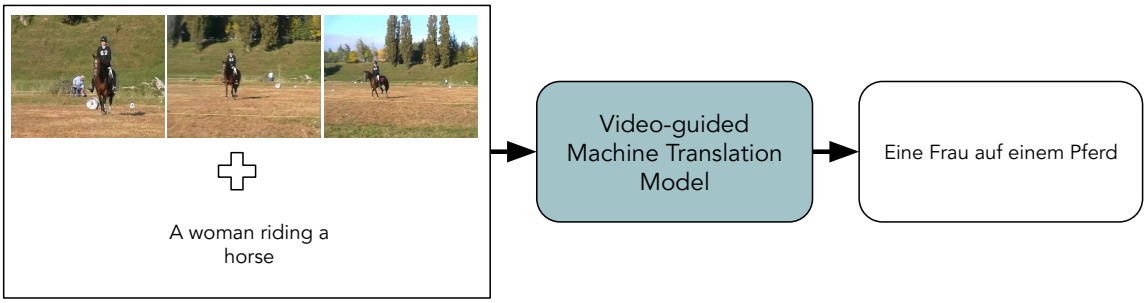

Figure 20: Video-guided Machine Translation.

by attending to source-language words and parts of an image independently by means of two separate attention mechanism.

MMT is also seen as two sub-tasks where learning to translate and learning visually grounded representations (Elliott & Kádár, 2017) in multi-task learning framework. Further, advanced multimodal compact bilinear pooling method (Delbrouck & Dupont, 2017a, 2017b) is also explored for MMT in which the outer product of two vectors to combine the attention features for the two modalities. Another model (Zhou et al., 2018d) used shared visual-language embedding and a translator for learning. This joint model leverages a visual attention grounding mechanism that links the visual semantics with the corresponding textual semantics. Due to presence of large multimodal data on the web, noisy image captions are also tried for MMT (Schamoni et al., 2018). A latent variable model (Calixto et al., 2018) is also used where the latent variable can be seen as a multi-modal stochastic embedding of an image and its description in a foreign language.

The MMT models were also explored in an adversarial setting. (Elliott, 2018) found that even in the presence of visual features from unrelated images there is no significant performance losses. Due to the success of unsupervised machine translation (Lample et al., 2018), there is also interest in extending it for unsupervised MMT (Su et al., 2018). Other studies (Caglayan et al., 2019) have reduced the criticism of MMT by showing that under the limited textual context, MMT models are capable of leveraging the visual input to generate better translations. While for multisource, (Libovický & Helcl, 2017) explored MMT from neural multi-source sequence-to-sequence learning.

2.10.2 Machine Translation with Video

The goal of video-guided machine translation (Wang et al., 2019) is to translate a source language description into the target language using the video information as additional spatio-temporal context. Figure 20 summarizes this approach where an video is accompanied with a English language description to be translated into the German description.

3. Datasets

There are wide range of datasets available for integration of vision and language research. They are the main driving force behind the fast advancement of this research area. Visual content associated with the textual content in these datasets differ from each other in size,

quality and the way they are collected. In this survey, we summarize the characteristics of these datasets and give an overview. However, we do not provide deeper analysis of them as in (Ferraro et al., 2015).

3.1 Image Description Generation

Many datasets were created in the past decade to address the challenge of image description generation. Some of the large datasets created earlier are about image captions and other image description datasets are small or medium scale. In the following, we cover those datasets which are popular and extensively used.

3.1.1 Flickr8K

The Flickr8K dataset (Hodosh et al., 2013) contain images from Flickr. The images in the dataset are selected through user queries for specific objects and actions using Amazon Mechanical Turk (AMT). Table 2 present splits of the dataset.

Split	Images	Captions per Image	Total Captions
Training	6,000	5	30,000
Validation	1,000	5	5,000
Test	1,000	5	5,000
Total	8,000	5	40,000

Table 2: Splits of "Flickr8K" image description dataset.

3.1.2 Flickr30K

The Flickr30K[1] dataset (Young et al., 2014) is an extended version of the Flickr8K dataset. It also contains images collected from Flickr and is annotated using AMT. Table 3 present splits of the dataset.

Split	Images	Captions per Image	Total Captions
Training	29,000	5	145,000
Validation	1,014	5	5,070
Test	1,000	5	5,000
Total	31,014	5	155,070

Table 3: Splits of "Flickr30K" image description dataset.

3.1.3 Flickr30K-Entities

The Flickr30K-entities[2] (Plummer et al., 2015) extends Flickr30K with manually annotated bounding boxes and entity mentions in captions. There are 276k bounding boxes for images and 513,644 entity mentions in the captions. Table 4 present splits of the dataset.

[1] https://shannon.cs.illinois.edu/DenotationGraph

[2] http://bryanplummer.com/Flickr30kEntities/

Split	Images	Captions per Image	Total Captions
Training	29,783	5	148,915
Validation	1,000	5	5,000
Test	1,000	5	5,000
Total	31,783	5	158,915

Table 4: Splits of "Flickr30K-Entities" image description dataset.

3.1.4 MSCOCO

The MSCOCO[3] dataset (Lin et al., 2014) contains natural images and are collected from the web. AMT is used to collect descriptions for images. It is a larger dataset when compared to Flickr datasets. Table 5 present splits of the dataset.

Split	Images	Captions per Image	Total Captions
Training	113,287	5	566,435
Validation	5,000	5	25,000
Test	5,000	5	25,000
Total	123,287	5	616,435

Table 5: Splits of "MSCOCO" image description dataset.

3.1.5 MSCOCO-Entities

MSCOCO-Entities[4] is a recently introduced dataset (Cornia et al., 2018) based on the original MSCOCO with more annotations. Table 6 present splits of the dataset.

Split	Images	Total Captions	Noun chunks	Noun chunks per caption	Unique Classes
Training	113,287	545,202	1,518,667	2.79	1,330
Validation	5,000	7,818	20,787	2.66	725
Test	5,000	7,797	20,596	2.64	730

Table 6: Splits of "MSCOCO-Entities" image description dataset.

3.1.6 STAIR

The STAIR[5] dataset consists of Japanese captions (Yoshikawa et al., 2017) for the 164,062 MSCOCO images. However, we have the same splits as of MSCOCO presented in Table 5. Original statistics of the dataset is provided in Table 7.

[3] http://cocodataset.org/#home
[4] https://github.com/aimagelab/show-control-and-tell
[5] http://captions.stair.center

Images	Total Captions	Vocabulary Size	Avg. number of chars
164,062	820,310	35,642	23.79

Table 7: Statistics of the "STAIR" image description dataset (Japanese).

3.1.7 Multi30K-CLID

The Multi30K-CLID[6] is designed for the Cross-Lingual Image Description (CLID) generation. Initially, Flickr30K dataset was extended with five German descriptions crowdsourced independently from their English versions. Hence, English-German pairs are considered as a comparable corpora. Table 8 present splits of the dataset.

Split	Images	**Language of the Captions**	
		English	German
Training	29,000	145,000	145,000
Validation	1,014	5,070	5,070
Testing	1,000	5,000	5,000

Table 8: Splits of "Multi30K-CLID" (2016) dataset.

Further, 2017 version[7] of the dataset is created with translations in French. Table 9 present splits of the dataset.

Split	Images	**Language of the Captions**		
		English	French	German
Training	29,000	145,000	145,000	145,000
Validation	1,014	5,070	5,070	5,070
Testing	1,000	5,000	5,000	5,000

Table 9: Splits of "Multi30K-CLID" (2017) dataset.

Similar to 2017 version, the 2018 version[8] of the dataset is added with additional translations in Czech language. Table 10 present more statistics about the dataset.

Split	Images	**Language of the Captions**			
		Czech	English	French	German
Training	29,000	145,000	145,000	145,000	145,000
Validation	1,014	5,070	5,070	5,070	5,070
Testing	1,071	5,355	5,355	5,355	5,355

Table 10: Splits of "Multi30K-CLID" (2018) dataset.

[6] https://www.statmt.org/wmt16/multimodal-task.html
[7] https://www.statmt.org/wmt17/multimodal-task.html
[8] http://www.statmt.org/wmt18/multimodal-task.html

3.1.8 Conceptual Captions

The conceptual captions[9] (Sharma et al., 2018) is a large scale dataset with more than 3.3M images paired with English language captions. The primary difference of this dataset in contrast with other image description datasets is that it is created automatically by crawling the web by extracting images and their corresponding captions from the alt-text of the HTML pages. Table 11 present splits of the dataset.

Split	Images	Captions
Training	3,318,333	3,318,333
Validation	15,840	15,840
Test	22,530	22,530

Table 11: Splits of "Conceptual Captions" image description dataset.

3.2 Video Description Generation

Similar to image description generation, several datasets were created to address the task of video description generation. In the following, we cover those datasets that are popular and extensively used. We denote hours → h, minutes → m and seconds → s.

3.2.1 Microsoft Video Description (MSVD)

It is an open domain dataset collected from Youtube clips[10] (Chen & Dolan, 2011) and annotated using AMT. The dataset is multilingual containing Chinese, English, German etc., human generated descriptions. On average, there are forty one single sentence descriptions per clip. Table 12 present the statistics about the dataset, while Table 13 presents its split.

Total Videos	Total Classes	Total Length	Avg. Length	Total Clips	Total Sentences	Total Words	Vocabulary Size
1,970	218	5.3 h	10 s	1,970	70,028	607,339	13,010

Table 12: Statistics of the "MSVD" dataset.

Split	Frames	Videos
Training	33,682	1200
Validation	3,275	100
Test	20,528	670

Table 13: Splits of "MSVD" dataset.

[9] https://ai.google.com/research/ConceptualCaptions/download
[10] http://www.cs.utexas.edu/users/ml/clamp/videoDescription

3.2.2 MPII Cooking

It consists of 65 different cooking activities[11] (Rohrbach et al., 2012). The dish preparation time ranges between 3 to 41 minutes. The 65 cooking activities include "wash hands", "put in bowl", etc. Table 14 present the statistics about the dataset, while Table 15 presents its split.

Number of Participants	Total Videos	Video Length	Total Length	Total Frames	Number of Activities	Total Dishes	Activity Annotations
12	44	3 to 41 m	8.0 h	881,755	65	14	5,609

Table 14: Statistics of the "MPII Cooking" dataset.

Split	Frames	Subjects
Training	1,071	10
Validation	-	-
Test	1,277	7

Table 15: Splits of "MPII Cooking" dataset.

3.2.3 YouCook

The YouCook[12] is a more complex dataset (Das et al., 2013), in contrast with MPII cooking and is annotated using AMT. Frames are annotated with objects belonging categories such as bowls, utensils etc., and actions. On average, eight descriptions per video is collected and is divided into six different cooking styles. Table 16 present the statistics about the dataset, while Table 17 presents its split.

Classes	Videos	Total Length	Sentences	Words	Vocabulary Size
6	88	2.3 h	2,688	42,457	2,711

Table 16: Statistics of the "YouCook" dataset.

Split	Videos
Training	49
Validation	-
Test	39

Table 17: Splits of "YouCook" dataset.

[11] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/mpii-cooking-activities-dataset

[12] http://web.eecs.umich.edu/~jjcorso/r/youcook

3.2.4 YouCook II

Similar to YouCook, the YouCook II[13] (Zhou et al., 2018b) also consists of cooking videos and is collected from YouTube. It has similar characteristics of open domain videos. Table 18 present the statistics about the dataset, while Table 19 presents its split.

Classes	Videos	Total Length	Avg. Length	Clips	Sentences	Vocabulary Size
89	2,000	176.0 h	316 s	15,400	15,400	2,600

Table 18: Statistics of the "YouCook II" dataset

Split	Videos
Training	1340
Validation	460
Test	200

Table 19: Splits of "YouCook II" dataset.

3.2.5 Textually Annotated Cooking Scenes (TACoS)

The TACoS dataset is a subset of MPII Composites[14] (Rohrbach et al., 2012) annotated with textual descriptions. It contains only those videos which contain activities such as manipulation of cooking ingredients. Around 26 cooking activities are collected with 127 videos. More details about the dataset is presented in Table 20 and Table 21. Generally, the dataset is split into 50% for training, 25% for validation, and 25% for testing.

Total Videos	Number of Clips	Descriptions per Video	Annotations after filtering	Cooking Tasks	Action Descriptions	Sentence Types
127	7,206	20	2,206	26	17,334 (tokens)	11,796

Table 20: Statistics of the "TACoS" dataset statistics - I

Total Words	Content Words (i.e., nouns, verbs, adjectives)	Total Verb Tokens	Total Verb Lemmas
146,771	75,210	28,292	435

Table 21: Statistics of the "TACoS" dataset statistics - II

3.2.6 TACoS-MultiLevel

The TACoS dataset is extended into TACoS-MultiLevel[15] (Rohrbach et al., 2014) by collecting three levels of descriptions constituting (i) Fifteen detailed description per video (ii)

[13] http://youcook2.eecs.umich.edu

[14] http://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos

[15] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/tacos-multi-level-corpus

Three to Five short descriptions and (iii) a single sentence description. More details about the dataset is presented in Table 22.

Total Videos	Total Clips	Total Video Length	Avg. Length	Number of Sentences	Total Words
185	14,105	27.1 h	360 s	52,593	2,000

Table 22: Statistics of the "TACoS-MultiLevel" dataset.

3.2.7 MPII Movie Description (MPII-MD)

The MPII-MD[16] (Rohrbach et al., 2015) contains clips extracted from the Hollywood movies and also their transcribed audio descriptions. Also, each clip is paired with one sentence that is extracted from the script of movie. Furthermore, transcribed audio is associated with spoken sentences by using time stamps. Misalignment between the audio and visual content is handled by leveraging manual annotation. Table 23 present the statistics of the dataset.

	Unique Movies	Before alignment Words	After alignment				
			Words	Sentences	Clips	Avg. Length	Total
Audio Desc.	55	346,557	332,846	37,272	37,266	4.1 s	42.5 h
Movie script	50	398,072	320,621	31,103	31,071	3.6 s	31.1 h
Total	94	744,629	653,467	68,375	68,337	3.9 s	73.6 h

Table 23: Statistics of the "MPII-MD" dataset.

For the task of video description, MPII-MD dataset is split as follows: 11 movies with associated scripts and audio descriptions (in total 22 alignments, 2 per movie) used as validation (8) and test set (14). The remaining 83 movies are used for training.

3.2.8 Montreal Video Annotation (M-VAD)

The M-VAD[17] (Torabi et al., 2015) is also created using 92 Movies. Table 24 present the statistics about the dataset, while Table 25 presents its split.

Type	Movies	Words	Paragraphs	Sentences	Avg. Length	Total
Un-filtered	92	531,778	52,683	59,415	6.3 s	91 h
Filtered	92	510,933	48,986	55,904	6.2 s	84.6 h

Table 24: Statistics of the "M-VAD" dataset.

[16] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/mpii-movie-description-dataset

[17] https://mila.quebec/en/publications/public-datasets/m-vad

Split	Videos
Training	38949
Validation	4888
Test	5149

Table 25: Splits of "M-VAD" dataset.

3.2.9 MSR Video to Text (MSR-VTT)

The MSR-VTT[18] (Xu et al., 2016) is created from open domain videos for video description generation. In addition to video content, dataset also contains audio information. Table 26 present the statistics about the dataset, while Table 27 presents its split.

Videos	Clips	Sentences	Words	Vocabulary	Duration
7,180	10,000	200,000	1,856,523	29,316	41.2 h

Table 26: Statistics of the "MSR-VTT" dataset.

Split	Video Clips
Training	6,513
Validation	497
Test	2,990

Table 27: Splits of "MSR-VTT" dataset.

3.2.10 Videos Titles in the Wild (VTW)

The VTW[19] dataset (Zeng et al., 2016) is created using video clips with an average of 90 seconds duration per clip and is described with one sentence to achieve video title generation. It also contains an augmented sentence with information not present in the video clip. Table 28 present the statistics about the dataset, while Table 29 presents its split.

Dataset	Sentences	Vocab	Sentences/Word	Nouns	Verbs	Adjective	Adverb
VTW-title	18,100	8,874	2.0	5,850	2,187	1,187	224
VTW-full	44,603	23,059	1.9	13,606	6,223	3,967	846

Table 28: Statistics of the "VTW" dataset.

3.2.11 ActivityNet Captions (ANetCap)

The ANetCap[20] (Krishna et al., 2017a) extends subset of videos from ActivityNet with descriptions. There are multiple descriptions for every video. It also contains multiple

[18] http://ms-multimedia-challenge.com/2017/dataset
[19] http://aliensunmin.github.io/project/video-language/index.html#VTW
[20] http://activity-net.org/challenges/2017/captioning.html

Split	Videos	Sentences
Training	14,100	14,100
Validation	2,000	2,000
Test	2,000	2,000

Table 29: Splits of "VTW" dataset.

events occurring at the same time. Table 30 present the statistics of the dataset, while Table 31 presents its split.

Videos	Total Video Hours	Avg. Video Length	Sentences	Avg. Sentence Length
20,000	849	180 s	100,000	13.48

Table 30: Statistics of the "ANetCap" dataset.

Split	Videos
Training	10,024
Validation	4,926
Test	5,044

Table 31: Splits of "ANetCap" dataset.

3.2.12 ActivityNet Entities (ANetEntities)

The ANetEntities[21] (Zhou et al., 2018a) dataset is based on the ANetCap (Section 3.2.11) dataset. It augments the ANetCap with 158k bounding box annotations, each grounded with a noun phrase. Table 32 present the dataset splits.

Split	Videos	Sentences	Objects	Bounding Boxes
Training	10,000	35,000	432	105,000
Validation	2,500	8,600	427	26,500
Test	2,500	8,500	421	26,100
Total	15,000	52,000	432	158,000

Table 32: Splits of the "ANetEntities" dataset.

3.3 Image Storytelling

Not many datasets were created to address the challenge of image storytelling. In the following, we cover all existing datasets that are used to address the task.

[21] https://github.com/facebookresearch/ActivityNet-Entities

3.3.1 New York City Storytelling (NYC-Storytelling)

The NYC-Storytelling[22] (Park & Kim, 2015) is created from the blogs where users post their travelogues. For evaluation, the dataset is split into 80%, 10%, and 10% for training, validation and testing respectively. Table 33 presents the statistics of the dataset.

Images	Blog posts
78,467	11,863

Table 33: Statistics of the "NYC-Storytelling" dataset.

3.3.2 Disneyland Storytelling

Similar to NYC-Storytelling, the Disneyland Storytelling dataset is also created from the travelogues. Same splits as of NYC-Storytelling (Section 3.3.1) is used for the evaluation. Table 34 presents the statistics of the dataset.

Images	Blog posts
60,545	7,717

Table 34: Statistics of the "Disneyland-Storytelling" dataset.

3.3.3 SIND

The Sequential Image Narrative Dataset (SIND) (Huang et al., 2016) is the first large dataset created for image storytelling. Natural language descriptions of the dataset is divided into three types: (i) Descriptions of Images-in-Isolation (DII), (ii) Descriptions of Images-in-Sequence (DIS), and (iii) Stories for Images-in-Sequence (SIS). Similar to other image storytelling datasets. This dataset is split into 80%, 10%, and 10% for training, validation and testing respectively. Table 35 presents the statistics of the dataset.

	Images	Flickr	(Text, Image)	Vocab
DII	-	-	151,800	13,800
DIS	-	-	151,800	5,000
SIS	-	-	252,900	18,200
Total	210,819	10,117	-	-

Table 35: Statistics of the "SIND" dataset.

3.3.4 VIST

Visual Storytelling Dataset (VIST)[23] is the second version of the SIND (Section 3.3.3). Table 36 present statistics of the dataset, while Table 37 present the splits of dataset.

[22] https://github.com/cesc-park/CRCN
[23] http://visionandlanguage.net/VIST

Images	Text Sequences
81,743	10,117

Table 36: Statistics of the "VIST" (SIND v.2) dataset.

Split	Images	Sentences
Training	40,155	200,775
Validation	4990	24,950
Test	5055	25,275

Table 37: Splits of "VIST" dataset.

3.4 Video Storytelling

Similar to image storytelling, two different datasets were created to address the challenge of video storytelling. In the following, we cover those datasets. However, both datasets are not publicly available.

3.4.1 VideoStory

The VideoStory (Gella et al., 2018) is a multi-sentence description dataset created from the social media videos. Table 38 present statistics of the dataset, while Table 39 present the splits of dataset.

Videos	Total Length	Clips	Sentences
20,000	396 h	123,000	123,000

Table 38: Statistics of the "VideoStory" dataset.

Split	Videos	Clips	Paragraphs	Words/paragraph
Training	17,098	80,598	17,098	61.76
Validation	999	13,796	2,997	59.88
Test	1,011	14,093	3,033	59.77
Test (Blind)	1,039	14,139	3,117	69.45
Total	20,147	122,626	26,245	62.23

Table 39: Splits of "VideoStory" dataset.

3.4.2 VideoStory-NUS

The VideoStory-NUS (Li et al., 2018) was collected from YouTube by querying four types of common and complex events such as birthday, camping, Christmas and wedding. Further, 105 videos with sufficient inter-event and intra-event variation are picked and annotated using AMT. Each video is annotated by at least 5 different workers and in total 529 stories were collected. Table 40 present statistics of the dataset, while Table 41 present the splits of dataset.

Domain	Videos	Avg. Video Length	Avg. Story Length	Avg. Sentence Length	Vocab Size
Open	105	12 m 35 s	162.6	12.1	4,045

Table 40: Statistics of the "VideoStory-NUS" dataset.

Split	Percentage (%)	Videos
Training	70	73
Validation	15	16
Test	15	16

Table 41: Splits of "VideoStory-NUS" dataset.

3.5 Image Question Answering

Several datasets were created in the past decade to address the challenge of image Question Answering. In the following, we cover those datasets that are extensively used for this task.

3.5.1 VQA v1.0

The VQA v1.0[24] dataset (Antol et al., 2015) contains open-ended questions about images. These questions target different areas of an image, including background details and underlying context. The answers are also open-ended and contain only a few words or a closed set of answers that can be provided in a multiple-choice format. Table 42 and Table 43 the dataset splits of the images with real and abstract scenes observed in the dataset respectively.

Dataset Split	Real Images	Questions per Image	Answers per Question	**Textual Annotations** Questions	Answers
Training	82,783	3	10	248,349	2,483,490
Validation	40,504	3	10	121,512	1,215,120
Test	81,434	3	10	244,302	-

Table 42: Splits of "VQA v1.0" dataset with real scenes.

Dataset Split	Abstract Scenes	Questions per Image	Answers per Question	**Textual Annotations** Questions	Answers
Training	20,000	3	10	60,000	600,000
Validation	10,000	3	10	30,000	300,000
Test	20,000	3	10	60,000	-

Table 43: Splits of "VQA v1.0" dataset with abstract scenes.

[24] `https://visualqa.org`

3.5.2 VQA v2.0

The VQA v2.0 dataset extends VQA v1.0 (Section 3.5.1) and has three parts: Balanced Real Images, Balanced Binary Abstract Scenes, and Abstract Scenes. Table 44 and Table 45 the dataset splits of the images with balanced real and binary abstract scenes observed in the dataset respectively. However, abstract scenes in VQA v2.0 is same as that of VQA v1.0.

Dataset Split	Real Images	Answers per Question	**Textual Annotations**		
			Questions	Answers	Complementary Pairs
Training	82,783	10	443,757	4,437,570	200,394
Validation	40,504	10	214,354	2,143,540	95,144
Test	81,434	10	447,793	-	-

Table 44: Splits of "VQA v2.0" with balanced real images.

Dataset Split	Binary Abstract Scenes	Answers per Question	**Textual Annotations**	
			Questions	Answers
Training	20,629	10	22,055	220,550
Validation	10,696	10	11,328	113,280
Test	-	-	-	-

Table 45: Splits of "VQA v2.0" with balanced binary abstract scenes.

3.5.3 OK-VQA

The OK-VQA[25] (Marino et al., 2019) dataset use subset of MSCOCO (Section 3.1.4). It is constructed with additional annotations such as questions, answers, knowledge category etc. Table 46 presents more details of the dataset, while Table 47 shows the split of dataset.

Total Images	Total Questions	Answers per Question	Unique Questions	Unique Answers	Unique Ques. Words	Total Categories	Average Ans. Length
14,031	14,055	5	12,591	14,454	7,178	10 + 1	1.3

Table 46: Statistics of the "OK-VQA" dataset.

Split	Percent (%)	Questions
Training	64	9,009
Validation	-	-
Test	36	5,046
Total	100	14,055

Table 47: Splits of "OK-VQA" dataset.

[25] https://okvqa.allenai.org

3.5.4 KVQA

The KVQA[26] (Shah et al., 2019b) was designed to emphasize on questions that require access to external knowledge. Table 48 presents more details of the dataset, while Table 47 shows the split of dataset. KVQA dataset provides five such splits instead of single split to get a mean score.

Total Images	Q&A Pairs	Unique Named Entities	Unique Answers	Avg. Ques. Len	Avg. Ans. Len	Avg. number of Questions per Image
24,602	183,007	18,880	19,571	10.14	1.64	7.44

Table 48: Statistics of the "KVQA" dataset.

Split	Percent (%)	Images	Q&A pairs
Training	70	17k	130k
Validation	20	5k	34k
Test	10	2k	19k

Table 49: Splits of "KVQA" dataset.

3.6 Video Question Answering

Similar to image question answering, several datasets were created to address the challenge of video question answering. In the following, we cover those datasets that are popular and extensively used.

3.6.1 MovieQA

The MovieQA[27] (Tapaswi et al., 2016) is used to evaluate story comprehension from both video and text in an automatic manner. The data set consists of almost 15,000 multiple choice question answers attained from over 400 movies having high diversity. Table 50 present the statistics and splits of dataset.

3.6.2 TVQA

The TVQA[28] (Lei et al., 2018) is created using six different TV shows such as Friends, The Big Bang Theory, How I Met Your Mother, House M.D., Grey's Anatomy, Castle. It consists of 460 hours of video and questions are designed to be compositional expecting the models to comprehend subtitles-based dialogue, and recognize relevant visual concepts. Table 51 presents the statistics of the dataset, while Table 52 shows the splits.

The testing data of TVQA is further split into two subsets named "test-public" containing 7,623 Q&A pairs and "test-reserved" consisting of 7,630 Q&A pairs. The *test-public* set is available in tvqa-leaderboard[29] whereas *test-reserved* is preserved for future use.

[26] http://malllabiisc.github.io/resources/kvqa
[27] http://movieqa.cs.toronto.edu/home
[28] http://tvqa.cs.unc.edu
[29] http://tvqa.cs.unc.edu/leaderboard.html

	Training	Validation	Test	Total
Movies with Plots and Subtitles				
Movies	269	56	83	408
QA pairs	9848	1958	3138	14944
Q words	9.3	9.3	9.5	9.3 ± 3.5
CA. words	5.7	5.4	5.4	5.6 ± 4.1
Movies with Video Clips				
Movies	93	21	26	140
QA pairs	4318	886	1258	6462
Video clips	4385	1098	1288	6771
Mean clip Length	201.0 s	198.5 s	211.4s	202.7 ± 216.2 s
Mean QA shots	45.6	49.0	46.6	46.3 ± 57.1

Table 50: Statistics & Splits for "MovieQA" dataset. The column 'Total' represents mean counts with standard deviations.

Video Clips	Video Clip Length	Q&A Pairs	Total Duration	Questions per Video Clip	Answers per Video Clip
21,793	60 to 90 s	152,545	460 h	7	5

Table 51: Statistics of the "TVQA" dataset.

Split	Percent (%)	Q&A pairs
Training	80	122,039
Validation	10	15,253
Test	10	15,253

Table 52: Splits of "TVQA" dataset.

The TVQA+[30] (Lei et al., 2019) is a subset and augmented version of the original TVQA dataset where the augmentation comes in the form of bounding boxes linking depicted objects to visual concepts in both questions and answers. Table 53 presents the splits of TVQA+ dataset.

Split	Q&As	Clips	Avg. Span Length (s)	Avg. Video Length (s)	Annotated Images	Bound. Boxes	Categories
Training	23,545	3,364	7.20	61.49	118,930	249,236	2,281
Validation	3,017	431	7.26	61.48	15,350	32,682	769
Test	2,821	403	7.18	61.48	14,188	28,908	680
Total	29,383	4,198	7.20	61.49	148,468	310,826	2,527

Table 53: Splits of "TVQA+" dataset.

[30] `http://tvqa.cs.unc.edu/download_tvqa_plus.html`

3.7 Image Dialog

3.7.1 VisDial

For the Image Dialog, there exists two versions of the dataset i.e., VisDial v0.9 and VisDial 1.0[31] (Das et al., 2017a). VisDial is created using the MSCOCO dataset. For VisDial v0.9, splits are divided only into the training and validation set. Table 54 and Table 55 present details about splits of VisDial v0.9 and VisDial v1.0 respectively.

Split	Images	Questions	Answers	Dialog Turns
Training	82,783	827,830	827,830	10
Validation	40,504	405,040	405,040	10
Test	-	-	-	-

Table 54: Splits of "VisDial v0.9" dataset.

Split	Images	Questions	Answers	Dialog Turns
Training	123,287	1,232,870	1,232,870	10
Validation	2,064	20,640	20,640	10
Test	8,000	80,000	80,000	1

Table 55: Splits of "VisDial v1.0" dataset.

3.7.2 CLEVR-Dialog

The CLEVR-Dialog[32] (Kottur et al., 2019) is developed for studying multi-round reasoning in visual dialog. The dialog grammar is grounded in the scene graphs of the CLEVR (Section 3.9.1) dataset originally developed for reasoning about images. Table 56 presents more details about the dataset, while Table 57 shows dataset splits.

CLEVR Images	Total Dialogs	Total Questions	Unique Questions	Unique Answers	Vocabulary Size	Dialog Turns	Mean Ques. Length
85k	425k	4.25M	73k	29	125	10	10.6

Table 56: Statistics of the "CLEVR-Dialog" dataset.

Split	Images	Q&A Pairs	Instances	Dialog Rounds
Training	70,000	3.5M	5	10
Validation	15,000	0.75M	5	10
Test	-	-	-	-

Table 57: Splits of "CLEVR-Dialog" dataset.

[31] https://visualdialog.org/data
[32] https://github.com/satwikkottur/clevr-dialog

3.8 Video Dialog

The Scene-Aware Dialog (AVSD)[33] (Alamri et al., 2019b) is created to address a challenge where the agent grounds its responses on the dynamic scene, the audio, and the history (previous rounds) of the dialog. Table 58 present more details of the dataset.

Split	Dialogs	Turns	Words
Training	7,985	123,480	1,163,969
Validation	1,863	14,680	138,314
Test	1,968	14,660	138,790

Table 58: Splits of "AVSD" dataset.

3.9 Image Reasoning

For image reasoning, both synthetic and real images datasets are developed. In the following, we present both synthetic and real image reasoning datasets.

3.9.1 Compositional Language and Elementary Visual Reasoning (CLEVR)

The CLEVR[34] (Johnson et al., 2017a) dataset is created with synthetic images rendered using Blender[35] toolkit. Table 59 presents the splits of dataset.

Split	Images	Questions	Unique Questions	Overlap with train
Training	70,000	699,989	608,607	-
Validation	15,000	149,991	140,448	17,338
Test	15,000	149,988	140,352	17,335
Total	100,000	999,968	853,554	-

Table 59: Splits of "CLEVR" dataset.

3.9.2 CLEVR-CoGenT

Modified version of CLEVR is Compositional Generalization Test (CLEVR-CoGenT)[34] (Johnson et al., 2017a). It is used to test models ability to find novel combinations of attributes at test-time. There are two types of conditions: Condition A and Condition B in this dataset, where based on the condition, the color of the geometrical shape can vary as show in the Table 60. Based on the conditions, the CLEVR-CoGenT dataset is divided as shown in the Table 61.

[33] https://video-dialog.com
[34] https://cs.stanford.edu/people/jcjohns/clevr
[35] https://www.blender.org

Geometrical Shape	Condition	Colors of Geometrical Shape
Cubes	A	gray, blue, brown, yellow
	B	red, green, purple, cyan
Cylinders	A	red, green, purple, cyan
	B	gray, blue, brown, yellow
Spheres	A	any color
	B	any color

Table 60: Conditions in "CLEVR-CoGenT" dataset.

Split	Condition	Images	Questions
Training	A	70,000	699,960
Validation	A	15,000	150,000
	B	15,000	149,991
Test	B	15,000	149,980
	B	15,000	149,992

Table 61: Splits of "CLEVR-CoGenT" dataset.

3.9.3 GQA

The GQA[36] (Hudson & Manning, 2019) is created to address shortcomings in earlier datasets. GQA consists of compositional questions over real-world images. Each image is associated with a scene graph of the image's objects, attributes and relations. Also, each question is associated with a structured representation of its semantics. Table 62 presents the statistics and splits of the dataset.

Images	Questions	Vocabulary Size	Training	Validation	Testing	Challenge
113,018	22,669,678	3,097	70%	10%	10%	10%

Table 62: Statistics & splits of the "GQA" dataset.

3.9.4 Relational and Analogical Visual rEasoNing (RAVEN)

The RAVEN[37] (Zhang et al., 2019) is created to perform relational and analogical visual reasoning. It is built by keeping in mind the Raven's Progressive Matrices (RPM) (Burke, 1958). Furthermore, it associates vision with structural, relational, and analogical reasoning in a hierarchical representation. The dataset is split into training, validation, and testing in the ratio 6:2:2 respectively. Table 63 presents the statistics of the dataset.

[36] https://cs.stanford.edu/people/dorarad/gqa

[37] http://wellyzhang.github.io/project/raven.html

Images	RPM Problems	Tree-structure per problem	Structural Labels	Rule Annotations	Avg. rules per problem
1,120,000	70,000	16	1,120,000	440, 000	6.29

Table 63: Statistics of the "RAVEN" dataset

3.10 Video Reasoning

The configurable visual question and answer (COG)[38] (Yang et al., 2018) is created to parallel experiments in humans and animals. Table 64 presents splits of the dataset.

Split	Total Examples	Examples per Task Family
Training	10,000320	227,280
Validation	500,016	11,364
Test	500,016	11,364

Table 64: Splits of "COG" dataset.

3.11 Image Referring Expression

For the image referring expression, both real and synthetic datasets are designed. In the following, we present the details of the datasets separately.

3.11.1 Real Images

For the real images, ImageCLEF and MSCOCO datasets are used to create the referring expression annotations. The per-object split is used to randomly divide the objects into training and testing sets. The RefCOCO[39] (Kazemzadeh et al., 2014), RefCOCO+ and RefCOCOg are created using MSCOCO images. For RefCOCO and RefCOCO+ , "People vs. Object" split evaluates images containing multiple people (Test A) and images containing multiple instances of all other objects (Test B). It is collected using a interactive setting (e.g., playing a game). Table 65 presents the statistics of the RefCOCO dataset whereas Table 66 shows the statistics of the RefCOCO+ dataset.

Images	Total Objects	Referring Expressions	Train/Test Splits
19,994	50,000	142,209	People vs. Object

Table 65: Statistics of the RefCOCO dataset

Unlike RefCOCO and RefCOCO+, RefCOCOg[40] (Mao et al., 2016) dataset presented in Table 67 is collected using a non-interactive setting and contains much longer sentences.

From ImageCLEF, RefClef[39] is collected and its statistics is presented in the Table 68.

[38] https://github.com/google/cog#datasets
[39] http://tamaraberg.com/referitgame
[40] https://github.com/lichengunc/refer

Images	Total Objects	Referring Expressions	Train/Test Splits
19,992	49,856	141,564	People vs. Object

Table 66: Statistics of the RefCOCO+ dataset

Images	Total Objects	Referring Expressions	Train/Test Splits
26,711	54,822	85,474	Per-Object

Table 67: Statistics of the RefCOCOg dataset

Real Images	Distinct Objects	Referring Expressions	Train/Test Splits
19,894	96,654	130,525	Per-Image split

Table 68: Statistics of the RefClef dataset.

Earlier mentioned datasets use single sentences for image referring expression. But, GuessWhat[41] (de Vries et al., 2017) dataset is created with a cooperative two-player guessing game to locate an unknown object in a image (collected from MSCOCO) by asking a sequence of questions. Hence, it creates a multiple sentences i.e., dialog for a given image to perform referring expression. Table 69 presents more details about the dataset. For evaluation, the dataset is split into 70% for training, 15% for validation, and 15% for testing.

	Images	Objects	Dialogues	Questions	Words	Vocab. Size
Full	66,537	134,073	155,280	821,889	3,986,192	11,465
Finished	65,112	125,349	144,434	732,081	3,540,497	10,985
Success	62,954	114,271	131,394	648,493	3,125,219	10,469

Table 69: Statistics of the "GuessWhat" dataset. The row 'Full' means all the dialogues are included, 'Finished' means all finished dialogues (successful and unsuccessful) are included, and 'Success' means only successful dialogues are included.

3.11.2 Synthetic Images

The CLEVR-Ref+[42] (Liu et al., 2019) is a synthetic dataset created to diagnose image reasoning with referring expressions. Table 70 present splits of the dataset.

3.12 Video Referring Expression

For performing Video Referring Expression, Cityscapes[43] dataset containing a diverse set of stereo video sequences recorded in street scenes is modified to have gaze information.

[41] https://guesswhat.ai

[42] https://cs.jhu.edu/~cxliu/2019/clevr-ref+

[43] https://www.cityscapes-dataset.com

Split	Images	Referring Expressions
Training	70,000	700,000
Validation	15,000	150,000
Test	15,000	150,000

Table 70: Splits of "CLEVR-Ref+" dataset.

ORGaze[44] (Balajee Vasudevan et al., 2018) contains object referring in videos with language and human gaze. More details of the dataset is presented in the Table 71.

Videos	Objects	Condition	Lighting	Annotations
5,000	30,000	Urban	Daytime	Bounding Boxes Gaze Recordings Language Expression

Table 71: Statistics of the "ORGaze" dataset

The authors split the cities in the training set of Cityscapes for training and validation while using all the cities in validation set for testing purposes. More concretely, the validation set is constructed by selecting one city (e.g., Zürich) from Cityscapes training set while leaving the rest of the cities as part of the training set. For constructing the testing set, the videos from all the cities in Cityscapes validation set (e.g., Frankfurt, Lindau, Münster) of Cityscapes are used.

Of the total 30,000 annotated objects, 80% has been used for *training* and the remaining 20% was reserved for model evaluation of the task.

3.13 Image Entailment

The image entailment is achieved using two different datasets. One dataset extends Natural Language Inference with visually-grounded Natural Language Inference (V-SNLI) (Vu et al., 2018). Another extends Flickr30K dataset (Section 3.1.2) into visual entailment dataset (SNLI-VE)[45] (Xie et al., 2019). Table 72 and Table 73 presents splits of the datasets.

Split	Entailment	Neutral	Contradiction
Training	182,167	181,515	181,938
Validation	3,329	3,235	3,278
Test	3,368	3,219	3,237
$\text{V-SNLI}_{\text{hard}}$ Test	1,058	1,068	1,135

Table 72: Splits of "V-SNLI" dataset.

[44] https://people.ee.ethz.ch/~arunv/ORGaze.html
[45] https://github.com/necla-ml/SNLI-VE

Split	Images	Entailment	Neutral	Contradiction	Vocab
Training	29,783	176,932	176,045	176,550	29,550
Validation	1000	5,959	5,960	5,939	6,576
Test	1000	5,973	5,964	5,964	6,592

Table 73: Splits of "SNLI-VE" dataset.

3.14 Image Generation

For image generation, existing image datasets are modified to accommodate image descriptions. Initially, Oxford-102[46] and Caltech-UCSD Birds (CUB)[47] datasets consisting of flower and bird images belonging to 102 and 200 classes respectively are expanded with image descriptions (Reed et al., 2016a). Table 74 and Table 75 presents splits of the datasets.

Split	Images	Captions per Image	Total Captions
Training	5,878	10	58,780
Validation	1,156	10	11,560
Test	1,155	10	11,550
Total	8,189	10	81,890

Table 74: Splits of "Oxford-102" dataset with image descriptions.

Split	Images	Captions per Image	Total Captions
Training	8,855	10	88,550
Validation	-	-	-
Test	2,933	10	29,330
Total	11,788	10	117,880

Table 75: Splits of "CUB" dataset with image descriptions.

Similarly, MSCOCO dataset (Section 3.1.4) is also used for reverse of description generation i.e., given a description, generate the image matching the description. We represent this dataset as MSCOCO-Gen. Table 76 present splits of dataset.

Split	Images	Captions per Image	Total Captions
Training	82,783	5	413,915
Validation	-	-	-
Test	40,504	5	202,520
Total	123,287	5	616,435

Table 76: Statistics of the "MSCOCO-Gen" dataset used for image generation.

[46] http://www.robots.ox.ac.uk/~vgg/data/flowers/102
[47] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

3.15 Video Generation

For the video generation there are no publicly available datasets. However, (Li et al., 2018b) has collected Text2Video dataset belonging to ten different categories of YouTube videos 400 each ranging between 10-400 seconds for language-to-video generation. The categories of videos are biking in snow, playing hockey, jogging, playing soccer ball, playing football, kite surfing, playing golf, swimming, sailing and water skiing. Table 77 shows the splits of the dataset.

Split	Videos
Training	2800
Validation	400
Test	800

Table 77: Splits of "Text2Video" dataset (Combines all categories).

3.16 Image-and-Language Navigation

For the image-and-language navigation, three different datasets were designed. In the following, we present the details of these datasets separately.

3.16.1 Room-2-Room (R2R)

The R2R dataset[48] (Anderson et al., 2018) consists of real images previously unseen, building-scale 3D environments. Table 78 presents splits of the dataset.

Split	Scenes	Instructions
Training	61	14,025
Validation (seen)	11	1,020
Validation (unseen)	11	2,349
Test	18	4,173

Table 78: Splits of "R2R" dataset.

3.16.2 ASKNAV

Similar to R2R dataset (Section 3.16.1), the ASKNAV[49] (Nguyen et al., 2019) is built on top of Matterport3D[50]. However, the objective remains different, where the agent queries the advisor when in confusion and make progress accordingly. It contains 10,800 panoramic views from 194,400 RGB-D images of 90 building-scale scenes. A data point in the dataset consists of a single starting viewpoint, but it has multiple goal viewpoints. Table 79 presents the splits of dataset.

[48] https://bringmeaspoon.org
[49] https://github.com/debadeepta/vnla
[50] https://niessner.github.io/Matterport

Split	Data points	Goals
Training	94,798	139,757
Validation (seen)	4,874	7,768
Validation (unseen)	5,005	8,245
Test (seen)	4,917	7,470
Test (unseen)	5,001	7,537

Table 79: Splits of "ASKNAV" dataset.

3.16.3 TOUCHDOWN

Extending from building environments, TOUCHDOWN[51] (Chen et al., 2019) dataset is designed for addressing tasks such as executing navigation instructions (Navigation Only) and resolving spatial descriptions (SDR) in the real-world environments. SDR is similar to the task of image referring expression (Section 2.6.1).

The environment includes 29,641 panoramas (360° Google Street View RGB images) and 61,319 edges from the New York City. Table 80 more details about the dataset, while Table 81 presents its splits.

Dataset	Dataset Size	Vocab. Size	Mean Text Length
TOUCHDOWN (Complete task)	9,326	5,625	108.0
Navigation Only	9,326	4,999	89.6
SDR Only	25,575	3,419	29.7

Table 80: Statistics of the "TOUCHDOWN" dataset. Vocabulary size and text length are computed by combining the training and validation sets.

Task	Split	Examples
Complete & Navigation Only	Training	6,526
	Validation	1,391
	Test	1,409
SDR Only	Training	17,880
	Validation	3,836
	Test	3,859

Table 81: Splits of "TOUCHDOWN" dataset.

3.17 Machine Translation with Image

The machine translation with image is achieved with Multi30K-MMT[52] dataset (Barrault et al., 2018) extended using Flickr30K dataset. Along with English, it contains human translated German, French and Czech sentences. Table 82 presents splits of the dataset.

[51] https://github.com/lil-lab/touchdown

[52] https://www.statmt.org/wmt18/multimodal-task.html

Split	Images	Captions
Training	29,000	29,000
Validation	1,014	1,014
Test	1,000	1,000

Table 82: Splits of "Multi30K-MMT" dataset for English, German, French and Czech.

3.18 Machine Translation with Video

The VATEX[53] (Wang et al., 2019) dataset is created for English and Chinese to perform machine translation with video and also multilingual video description generation. Table 82 present more details about the dataset.

Split	Videos	Action Label
Training	25,991	✓
Validation	3,000	✓
Public Test	6,000	-
Secret Test	6,278	-

Table 83: Splits of "VATEX" dataset. The secret test set denote the human-annotated captions holdout for challenge.

3.19 Miscellaneous

In this section, we present those datasets which are not task-specific. However, they have either contributed indirectly to the tasks mentioned in this survey or pushing the integration of language and vision research to new domains.

3.19.1 Visual Genome

To comprehend interactions and relationships between objects observed in an image, Visual Genome[54] (Krishna et al., 2017b) dataset is created. Annotation such as objects, attributes, and relationships within each image is collected in the dataset. However, dataset has also become the foundation of many other vision and language integration tasks. Table 84 present more details about the dataset.

Total Images	Descriptions per Image	Total Objects	Object Categories	Attributes Categories	Relationship Categories	Question Answers
108,000	50	4,102,818	76,340	15,626	47	1,773,258

Table 84: Statistics of the Visual Genome dataset

[53] http://vatex.org/main/index.html
[54] https://visualgenome.org

3.19.2 How2

The How2[55] (Sanabria et al., 2018) dataset is a multilingual and multimodal collection which consists of instructional videos with English (En) language subtitles paired with crowd-sourced Portuguese (Pt) translations. Table 85 presents the splits of dataset.

Name	Dataset Split	Total Videos	Total Hours	Clips / Sentences	Per Clip Statistics	Tokens En	Tokens Pt
300 h	Training	13,168	298.2	184,949	5.8 s & 20 words	3.8M(43k)	3.6M(60k)
	Validation	150	3.2	2,022	5.8 s & 20 words	-	-
	Test	175	3.7	2,305	5.8 s & 20 words	-	-
	Held-out	169	3.0	2,021	5.4 s & 20 words	-	-
2000 h	Training	73,993	1,766.6	-	-	-	-
	Validation	2,965	71.3	-	-	-	-
	Test	2,156	51.7	-	-	-	-

Table 85: Splits of "How2" dataset. The *300 h* is a subset of the full set i.e., *2000 h*. The Held-out represents the set that has been reserved for future evaluations. The information in brackets indicate unique tokens.

3.19.3 Berkeley Deep Drive eXplanation (BDD-X)

The BDD-X[56] (Kim et al., 2018) dataset is created to provide explanations to the autonomous driving conditions. It can be seen analogous to the vision-and-language navigation task. BDD-X is built on top of the Berkeley Deep Drive (BDD)[57] dataset, which contains dashboard camera videos of approximately 40 seconds in duration. This dataset is composed of over 77 hours of driving videos and are taken in diverse driving conditions, e.g., day/night, highway/city/countryside, summer/winter etc. On average each video is 40 seconds in duration and contains about 3-4 actions, e.g., speeding up, slowing down, turning right etc., all of which are annotated with a description and an explanation. Table 86 present the statistics of the dataset, while Table 87 shows the dataset split.

Videos	Frames	Condition	Lighting	Hours	Annotations	Actions per video
6,984	8,400,000	Urban	Day/Night	77 h	26,228	3 to 4

Table 86: Statistics of the "BDD-X" dataset.

4. Representation

In this section, we briefly present the neural network architectures used to build vector representations of both vision and language.

[55] https://srvk.github.io/how2-dataset

[56] https://github.com/JinkyuKimUCB/BDD-X-dataset

[57] https://bdd-data.berkeley.edu

Split	Videos
Training	5,588
Validation	698
Test	698

Table 87: Splits of "BDD-X" dataset.

4.1 Vision

With the advent of deep learning (LeCun et al., 2015), best way to extract vector representation from the visual data is by leveraging automatic feature extraction methods. In the following, we present different neural network architectures that are used to generate representations for both images and videos.

4.1.1 Image Representation

Convolutional neural networks (CNN) (LeCun et al., 1995) have become de facto standard for generating representation for images with an end-to-end trainable framework. There are several variations of CNNs proposed to learn image features with supervised or self-supervised (Jing & Tian, 2019) techniques. Most of these techniques are aimed to learn transferable general image features by leveraging tasks such as image classification, detection, semantic segmentation, and action recognition. Usually, most preferred transferrable global image representations are learned with deep CNN architectures such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan & Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), Inception-v3 (Szegedy et al., 2015), Residual Nets (ResNet) (He et al., 2016) and Dense Nets (Huang et al., 2017) using large datasets such as ImageNet (Deng et al., 2009), MSCOCO (Section 3.1.4)and Visual Genome (Section 3.19.1). However, some language and vision integration tasks prefer to learn global image features during task-specific training as opposed to pretrained representations.

For learning local features i.e., features of objects present in the images represented with bounding boxes, the preferred choice is to utilize region specific CNN architecture such as Region-based CNN (R-CNN) (Ren et al., 2015c).

4.1.2 Video Representation

Videos extend images present in the 3D channel into 4D. Generally, visual data observed in videos extracted in the form of screenshots leverage same techniques as for image local and global representation. However, in addition to it, spatio-temporal features are also developed with general video analysis such as C3D (Tran et al., 2014) or from action recognition dataset i.e., Kinetics action recognition (Kay et al., 2017) to build R3D or I3D features (Carreira & Zisserman, 2017) using different CNN architectures.

4.2 Language

In most cases, language is represented either with bag-of-words or with sentence representations. For words in a sentence, initializations are generally done with pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014). Further, to represent variable-

length text, sequence learning techniques are applied such as recurrent neural networks and its variations like unidirectional (LSTM) or bidirectional (BiLSTM) long short-term memory (Hochreiter & Schmidhuber, 1997) and unidirectional (GRU) or bidirectional (BiGRU) gated recurrent units (Chung et al., 2014). Recently, to provide parallelization in sequential training, self-attention approaches such as Transformer (Vaswani et al., 2017) are designed.

4.3 Vision and Language

In earlier sections, we saw different neural network architectures used to represent both vision and language separately. In this section, we present the architectures used by state-of-the-art methods for combining representations of language and vision to address different tasks. We also highlight the gradient descent optimization algorithms such as stochastic gradient descent (SGD) (Bottou, 2010), ADAM (Kingma & Ba, 2014), RMSprop (Tieleman & Hinton, 2012) used by these methods for training. Furthermore, we check if the methods leverage reinforcement learning (RL). We present details for only eight tasks by excluding visual description generation (Aafaq et al., 2018; Hossain et al., 2019) and question answering (Wu et al., 2017b) as they are extensively reviewed in the recent surveys.

4.3.1 Visual Storytelling

As discussed in the Section 2.2, the visual storytelling generates a story like narrative given an image or a video. In this section, we present the representation used for both vision i.e., image or video and language by major visual storytelling architectures which integrate them. In Table 88, different architectures (refer to Combined column) built for image storytelling is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Kiros et al., 2014a)	AlexNet	LM	MLBL	-	✗
(Karpathy & Fei-Fei, 2015)	VGG	RNN	NeuralTalk	RMSprop	✗
(Vinyals et al., 2015)	GoogLeNet	LSTM	NIC	SGD	✗
(Park & Kim, 2015)	VGG	RNN	CRCN	RMSprop	✗
(Huang et al., 2016)	VGG	GRU	Story-Flat	-	✗
(Krause et al., 2017)	VGG	LSTM	HierarchicalRNN	ADAM	✗
(Liu et al., 2017)	VGG	LSTM	BARNN	-	✗
(Wang et al., 2018a)	VGG	LSTM	GAN	ADAM	✓
(Wang et al., 2018a)	ResNet-152	GRU	AREL	ADAM	✓

Table 88: Major Image Storytelling Architectures.

While in the Table 89, architectures (refer to Combined column) built for video storytelling is presented.

4.3.2 Visual Dialog

As discussed in the Section 2.4, the visual dialog system hold a dialog with humans using natural language about a given image or a video. In this section, we present the representation used for both vision i.e., image or video and language by the major visual dialog

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Yu et al., 2016)	C3D	VGG	GRU	H-RNN	RMSProp	✗
(Gella et al., 2018)	R3D	ResNet-101	GRU	seq-seq+context	ADAM	✗
(Li et al., 2018)	-	ResNet-101	GRU	ResBRNN	ADAM	✓

Table 89: Major Video Storytelling Architectures. TP-Temporal Pyramid

architectures which integrate them. In Table 90, different architectures (refer to Combined column) built for image dialog is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Das et al., 2017a)	VGG	LSTM	MemoryNetwork	ADAM	✗
(Lu et al., 2017)	VGG	LSTM	HCIAE-NP-ATT	ADAM	✗
(Seo et al., 2017)	VGG	LSTM	AMEM	ADAM	✗
(Jain et al., 2018)	VGG	LSTM	SF	ADAM	✗
(Kottur et al., 2018)	ResNet-152	LSTM	CorefNMN	-	✗
(Wu et al., 2018)	VGG	LSTM	CoAtt-GAN	ADAM	✓
(Niu et al., 2019)	ResNet-152	LSTM	RvA	ADAM	✗
(Zheng et al., 2019)	VGG	LSTM	GNN	ADAM	✗
(Guo et al., 2019)	ResNet-101	LSTM	Synergistic	ADAM	✗

Table 90: Major Image Dialog Architectures (Discriminative and Generative).

While in the Table 91, architectures (refer to Combined column) built for video dialog is presented.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Hori et al., 2019)	I3D	VGG	LSTM	MultimodalAtt	ADAM	✗
(Schwartz et al., 2019)	I3D	VGG	LSTM	i3d-rgb-spatial-10	ADAM	✗

Table 91: Major Video Dialog Architectures.

4.3.3 Visual Reasoning

As discussed in the Section 2.5, the visual reasoning system answer sophisticated queries by reasoning about an image or a video. In this section, we present the representation used for both vision i.e., image or video and language by the major visual reasoning architectures which integrate them. In Table 92, different architectures (refer to Combined column) built for image reasoning is presented.

While in the Table 93, architectures (refer to Combined column) built for video reasoning is presented.

4.3.4 Visual Referring Expression

As discussed in the Section 2.6, the visual referring expression system should use referring expressions to unambiguously identify or indicate particular objects in an image or a video. In this section, we present the representation used for both vision i.e., image or video

Approach	Image	Language	Combined	Optimizer	RL
(Johnson et al., 2017a)	ResNet-101	LSTM	SA+MLP	ADAM	✗
(Hu et al., 2017)	VGG	LSTM	N2NMN	ADAM	✓
(Johnson et al., 2017b)	ResNet-101	LSTM	PGEE	ADAM	✓
(Santoro et al., 2017)	Custom	LSTM	RN	ADAM	✗
(Cao et al., 2018)	ResNet-101	BiLSTM	ACMN	ADAM	✗
(Perez et al., 2018)	ResNet-101	GRU	FiLM	ADAM	✗
(Hudson & Manning, 2018)	ResNet-101	BiLSTM	MAC	ADAM	✗
(Mascharka et al., 2018)	ResNet-101	-	TbD	ADAM	✗
(Haurilet et al., 2019)	ResNet-152	LSTM	FinalDestGraph	ADAM	✗
(Hu et al., 2019)	ResNet-101	LSTM	LCGN	ADAM	✗
(Mao et al., 2019)	ResNet-34	BiGRU	NS-CL	-	✓

Table 92: Major Image Reasoning Architectures. Custom - Own CNN architecture.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Yang et al., 2018)	-	Custom	LSTM	WorkMemory	ADAM	✗
(Haurilet et al., 2019)	-	ResNet-152	LSTM	FinalDestGraph	ADAM	✗

Table 93: Major Video Reasoning Architectures.

and language by the major visual referring expression architectures which integrate them. In Table 94, different architectures (refer to Combined column) built for image referring expression comprehension is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Mao et al., 2016)	VGG	LSTM	MMI	SGD	✗
(Nagaraja et al., 2016)	VGG	LSTM	Neg. Bag	SGD	✗
(Yu et al., 2016)	VGG	LSTM	Context	-	✗
(Luo & Shakhnarovich, 2017)	VGG	BiLSTM	CG	ADAM	✗
(Liu et al., 2017)	VGG	LSTM	Combined	ADAM	✗
(Hu et al., 2017b)	VGG	LSTM	CMN	-	✗
(Yu et al., 2017a)	VGG	LSTM	Reinforcer	ADAM	✓
(Zhang et al., 2018b)	VGG	BiLSTM	VarContext	SGD	✓
(Deng et al., 2018)	VGG	LSTM	AccumulateAtt	SGD	✗
(Zhuang et al., 2018)	VGG	LSTM	ParallelAtt	ADAM	✗
(Yu et al., 2018)	ResNet-101	BiLSTM	MAttNet	-	✗
(Hong et al., 2019)	ResNet-101	BiLSTM	RVG-Tree	ADAM	✗
(Yang et al., 2019)	ResNet-101	BiLSTM	CMRIN	ADAM	✗

Table 94: Major Image Referring Expression Comprehension Architectures.

While in the Table 95, architectures (refer to Combined column) built for video referring expression comprehension is presented.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Balajee Vasudevan et al., 2018)	-	VGG	LSTM	WithGaze	-	✗

Table 95: Major Video Referring Expression Comprehension Architectures.

4.3.5 Visual Entailment

As discussed in the Section 2.7, the visual entailment system should predict whether the image or a video semantically entails the text. In this section, we present the representation used for both vision i.e., image or video and language by the major visual entailment architectures which integrate them. In Table 94, different architectures (refer to Combined column) built for image entailment is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Vu et al., 2018)	VGG	BiLSTM	V-BiMPM	ADAM	✗
(Xie et al., 2019)	ResNet-101	GRU	EVE-Image	ADAM	✗

Table 96: Major Image Entailment Architectures.

4.3.6 Language-to-Vision Generation

As discussed in the Section 2.8, the language-to-vision generation system should generate a image or a video given global description of it. In this section, we present the representation used for both vision i.e., image or video and language by the major language-to-vision architectures which integrate them. In Table 97, different architectures (refer to Combined column) built for language-to-image is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Reed et al., 2016a)	-	char-CNN-RNN	GAN-INT-CLS	ADAM	✗
(Reed et al., 2016b)	-	char-CNN-GRU	GAWWN	ADAM	✗
(Zhang et al., 2017)	-	-	StackGAN	ADAM	✗
(Xu et al., 2018)	Inception-v3	BiLSTM	AttGAN	-	✗
(Qiao et al., 2019)	-	BiLSTM	MirrorGAN	-	✗

Table 97: Major Language-to-Image Generation Architectures.

While in the Table 98, architectures (refer to Combined column) built for language-to-video is presented.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Li et al., 2018b)	MotionFeatures	-	LSTM	T2V	ADAM	✗

Table 98: Major Language-to-Video Generation Architectures.

4.3.7 Vision-and-Language Navigation

As discussed in the Section 2.9, the vision-and-language navigation system should reach its destination using both visual and language instructions. In this section, we present the representation used for both vision i.e., image or video and language by the major vision-and-language navigation architectures which integrate them. In Table 99, different architectures (refer to Combined column) built for image-and-language navigation is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Anderson et al., 2018)	ResNet-152	LSTM	Seq-to-Seq	ADAM	✗
(Wang et al., 2018)	ResNet-152	LSTM	RPA	-	✓
(Fried et al., 2018)	ResNet-152	LSTM	Speaker-Follower	-	✓
(Wang et al., 2018c)	ResNet-152	LSTM	RCM	ADAM	✓
(Ma et al., 2019a)	ResNet-152	LSTM	Self-Monitoring	ADAM	✗
(Tan et al., 2019)	ResNet-152	LSTM	BackTranslation	RMSprop	✓
(Ke et al., 2019)	-	LSTM	TacticalRewind	-	✗

Table 99: Major Image-and-Language Navigation Architectures.

4.3.8 Multimodal Machine Translation

As discussed in the Section 2.10, the multimodal machine translation system should translate source natural language description of visual content using both vision and language information. In this section, we present the representation used for both vision i.e., image or video and language by the major multimodal machine translation architectures which integrate them. In Table 100, different architectures (refer to Combined column) built for machine translation with image is presented.

Approach	Image	Language	Combined	Optimizer	RL
(Calixto et al., 2017)	ResNet-50	BiGRU	DoubleAtt	Adadelta	✗
(Calixto & Liu, 2017)	VGG	BiGRU	GVF	Adadelta	✗
(Elliott & Kádár, 2017)	Inception-V3*	BiGRU	Imagination	ADAM	✗
(Caglayan et al., 2017)	ResNet-50	BiGRU	Lium-cvc-ensemble	ADAM	✗
(Calixto et al., 2018)	ResNet-50	BiGRU	$VMMT_F$	ADAM	✗
(Helcl et al., 2018)	ResNet-50	LSTM	CUNI-ensemble	ADAM	✗

Table 100: Major Machine Translation with Image Architectures. * - compares with ResNet-50 and VGG also.

While in the Table 101, architectures (refer to Combined column) built for machine translation with video is presented.

Approach	Video	Frame	Language	Combined	Optimizer	RL
(Wang et al., 2019)	I3D	-	LSTM	NMT+LSTM VI	ADAM	✗

Table 101: Major Machine Translation with Video Architectures.

5. Evaluation Measures

We segregate the evaluation measures into three different categories where the first set of measures used are common across several tasks, the second category represent task-specific measures and the third category denote human evaluation. In the following, we present details for each of them separately.

5.1 Common Measures

We divide common measures further into two different categories: (i) Language Metrics and (ii) Retrieval Metrics.

5.1.1 Language Metrics

We label those metrics as "language metrics" which evaluate the machine generated text based on reference text using word overlaps.

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) was proposed for machine translation to compare machine generated output with the human ground truth. BLEU calculates the overlap between predicted unigrams (BLEU-1 (B-1)) or n-grams (BLEU-2 (B-2), BLEU-3 (B-3) and BLEU-2 (B-4)) from the set of candidate reference sentences. Tasks such as Visual Caption Generation (Section 2.1), Visual Storytelling (Section 2.2), Video Dialog (Section 2.4.2) and Multimodal Machine Translation (Section 2.10) use BLEU as its evaluation measure. To achieve high BLEU score, generated descriptions should match the human ground truth words as well as their order. Maximum BLEU score is one for an exact match between generated and reference sentence.

Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee & Lavie, 2005) overcome the issues in BLEU such as exact word match. METEOR performs semantic matching by leveraging WordNet. METEOR used WordNet to match words at various levels with synonymy and the paraphrase matching. METEOR score is computed using the alignment between the machine generated output and the reference sentences. Initially, set of unigrams from generated and reference sentence is used to perform alignment is done. If there are multiple options available for alignments between the generated and reference sentence, the alignment setting with less number of comparisons is preferred. After finalizing the alignment process, METEOR score is calculated. Tasks such as Visual Caption Generation (Section 2.1), Visual Storytelling (Section 2.2), Video Dialog (Section 2.4.2) and Multimodal Machine Translation (Section 2.10) use METEOR as its evaluation measure.

Recall Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) was designed to evaluate textual summaries. As opposed to BLEU, which concentrate on n-gram precision, ROUGE calculate recall score of the generated sentences corresponding to the reference sentences. Most prominent ROUGE variant used in ROUGE-L which is based on longest common subsequence and is used by Visual Caption Generation (Section 2.1) for evaluation. Other variants include ROUGE-W (Weighted Longest Common Sub-sequence) and ROUGE-S (Skip-Bigram Co-Occurrences Statistics). The advantage of ROUGE-L over BLEU and METEOR is that it checks for the insequence within a sentence. Moreover,

mentioning the n-gram length (e.g., BLEU) is also not required as it is automatically incorporated.

Consensus based Image Description Evaluation (CIDEr) (Vedantam et al., 2015) was initially designed for image caption generation (Section 2.1.1) evaluation. However, it is also adopted for Video Caption Generation (Section 2.1.2), Visual Storytelling (Section 2.2) and Video Dialog (Section 2.4.2). CIDEr evaluates the consensus between a generated and reference sentences by performing different language pruning techniques such as stemming and building a set of n-grams. Finally, n-grams that are common among the reference sentences of all visual data is given lower weight, as they are supposed to be less informative about the visual content, and biased towards textual content of the sentences. The weight for each n-gram is computed using Term Frequency Inverse Document Frequency (TF-IDF), where TF puts higher weightage on frequently occurring n-grams in the reference sentence of the visual content, whereas IDF puts lower weightage on commonly appearing n-grams across the whole dataset.

To remove the mismatch between human evaluation and CIDEr scores, a variant of CIDEr i.e., CIDEr-D is proposed. It adopts small variations such as no stemming and ensure the words of high confidence are not repeated in a sentence by introducing a Gaussian penalty over length differences between the generated and reference sentences. As in the CIDEr, it produces high score even if the sentence does not make sense.

Semantic Propositional Image Captioning Evaluation (SPICE) (Anderson et al., 2016) was initially designed for image caption generation (Section 2.1.1) evaluation and is also adopted for Video Caption Generation (Section 2.1.2). SPICE measures the similarity between the scene graph tuples parsed from the generated sentences and the human created ground truth reference sentences. The scene graph encodes objects, relationships and their relationships through a dependency parsing. Hence, it makes SPICE heavily dependent on parsing which can be prone to errors. Similar to METEOR, SPICE uses WordNet to find and treat synonyms as positive matches to compute F1-score between the tuples of generated sentences and the ground truth.

5.1.2 Retrieval Metrics

We label those metrics as "retrieval metrics" which evaluate the machine generated text based on standard information retrieval (Manning et al., 2010) metrics.

Recall@k (R@k) goal is to evaluate the number of relevant ground truth sentences retrieved in the Top-k (e.g., Top-1, Top-5 etc.) candidates. Tasks such as Visual Caption Generation (Section 2.1), Visual Storytelling (Section 2.2) and Image Dialog (Section 2.4.1) use it for evaluation. Higher R@k indicates better performance.

Median Rank (MedRank) find the median rank value of the retrieved ground truth. Tasks such as Visual Storytelling (Section 2.2) use this metric for evaluation. Lower MedRank value indicates better performance.

Mean Reciprocal Rank (MRR) is a binary measure, where the rank of the highest ranking relevant document for a query is used to calculate the reciprocal rank averaged over

all queries. Tasks such as Image Dialog (Section 2.4.1) use it for evaluation. Higher MRR indicates better performance.

Mean Rank (Mean) refer to the mean rank achieved in retrieving the relevant sentence. Tasks such as Image Dialog (Section 2.4.1) use it for evaluation. Lower value is better for Mean.

Normalized Discounted Cumulative Gain (NDCG) is a variant of the Discounted Cumulative Gain (DCG) (Järvelin & Kekäläinen, 2000), NDCG is a cumulative, multilevel measure of ranking quality that is usually truncated at a particular rank level. Tasks such as Image Dialog (Section 2.4.1) use it for evaluation.

5.2 Task-specific Metrics

The task-specific metrics denote those measures which are used specific to a task. In the following, we represent those tasks which use such measures in addition or independent of common measures.

5.2.1 Visual Reasoning

There are evaluation measures which are used specifically for Image Reasoning (Section 2.5.1) (e.g., CLEVR dataset) and is discussed in detail in the following.

Querying attribute (QA) use questions to ask about an attribute of a particular object.

Compare Attribute (CA) use compare questions to ask whether two objects have the same value for some attribute.

Compare Numbers (CN) use comparison questions to ask which of two object sets is larger.

Count ask counting questions to find number of objects fulfilling some conditions.

Exist ask existence questions whether a certain type of object is present.

Similarly, for the Video Reasoning (Section 2.5.1) (e.g., COG dataset) there are different metrics used for evaluation. The evaluation measures are based on account changes of the scene in three different query types:

Pointing (Point) use questions to ask about pointing to a certain object.

Yes/No use questions about binary decision.

Conditional (Condit) are questions based on objects that needs to fulfill certain conditions.

Attribute-related (Atts) are questions about certain attributes.

5.2.2 Language-to-Vision Generation

There are evaluation measures which are used specifically for language-to-image generation (Section 2.8.1) and is discussed in detail in the following.

Inception Score (IS) (Salimans et al., 2016) was initially proposed to compare the quality of images generated by GAN models. A pretrained Inception-v3 model (Szegedy et al., 2016) is applied to the generated image to get the conditional label distribution with low entropy. Similar idea is applied for the generated images on the given text descriptions for automatic evaluation. Higher scores are better for IS.

Fréchet Inception distance (FID) (Heusel et al., 2017) is supposed to improve on the IS by comparing the statistics of generated samples to original samples, instead of evaluating generated samples in an isolated manner. It also depends on the Inception-v3 model. Especially, pool3 layer of the Inception-v3 is used for generated and original samples for comparison. Lower FID is better as it corresponds to more similar generated and original samples.

R-precision is inspired from the ranking retrieval results. It is used as a complementary evaluation metric for the language-to-image generation. Especially, generated images are used to query their corresponding natural language descriptions to find how many relevant descriptions are retrieved.

5.2.3 Vision-and-Language Navigation

There are evaluation measures (Anderson et al., 2018) which are used specifically for image-and-language navigation (Section 2.9.1) and is discussed in detail in the following.

Path Length (PL) is seen as trajectory length where it is the total length of the executed path.

Navigation Error (NE) is a shortest path distance in the navigation graph i.e., measuring the average distance between the end-location predicted by the follower agent and the true route's end-location.

Success Rate (SR) is the percentage of predicted end-locations within 3m of the true location.

Oracle Success Rate (OSR) measures the success rate at the closest point to the goal that the agent has visited along the trajectory.

Success Path Length (SPL) trades-off Success Rate against Path Length (PL) i.e., success rate weighted by inverse PL.

5.3 Human Evaluation

Earlier mentioned metrics provide only quantitative measures for evaluating different tasks. However, due to lack of high correlation between machine generated textual or visual data with the human provided ground truth, most of the tasks require human evaluations to judge the quality of the content. Especially, tasks such as Visual Caption Generation (Section 2.1), Visual Storytelling (Section 2.2), language-to-image generation (Section 2.8.1) employ crowd workers to evaluate quality of the content generated.

To perform evaluation, based on the task various kinds of instructions are given to human evaluators. Most of the tasks are interested in finding relevance of the output to input. Also, they evaluate to decide the preference of the methods based on the generated output.

6. State-of-the-Art Results

In this section, we summarize the performance of various methods for eight tasks[58] presented in this survey. Results are shown for each dataset separately and methods are ordered chronologically. By grouping the results based on dataset, we can infer the difficulty level of datasets by comparing the intra-dataset scores for same methods. Moreover, for different variations of the same model, we only report their best results. In addition, we tried to represent all measures used for evaluations, however, due to space limitations some tables are presented only with prominent metrics.

6.1 Visual Storytelling Results

In the following, we present the results achieved with various methods for image and video storytelling separately.

6.1.1 Image Storytelling

To evaluate image storytelling, we use the datasets presented in the Section 3.3. Both language and retrieval metrics are used as evaluation measures for comparison of various image storytelling methods. Table 102, Table 103, Table 104 and Table 105 present the benchmark results on different datasets separately.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
MLBL (Kiros et al., 2014a)	0.01	2.6	5.29	1.19	4.52	100.5
NeuralTalk (Karpathy & Fei-Fei, 2015)	0.00	0.5	1.34	0.48	2.86	120.5
NIC (Vinyals et al., 2015)	0.10	9.1	5.73	0.95	7.38	88.5
CRCN (Park & Kim, 2015)	**2.08**	30.9	7.69	11.67	31.19	14.00
Story-Flat (Huang et al., 2016)	-	-	7.37	-	-	-
HierarchialRNN (Krause et al., 2017)	-	-	6.07	-	-	-
BARNN (Liu et al., 2017)	-	**41.6**	-	**29.37**	**45.43**	**8**
Adversarial (Wang et al., 2018a)	-	-	**8.39**	-	-	-

Table 102: Language and Retrieval Metrics of different methods on the "NYC-Storytelling" (Section 3.3.1) dataset.

6.1.2 Video Storytelling

To evaluate video storytelling, two different datasets presented in the Section 3.4 are used. Both language and retrieval metrics are used as evaluation measures for comparing different methods. Table 106 and Table 107 present the benchmark results on each of these datasets separately.

6.2 Visual Dialog Results

In the following, we present the results achieved with various methods for image and video dialog separately.

[58] We skip visual description generation (Aafaq et al., 2018; Hossain et al., 2019) and question answering (Wu et al., 2017b) as it is extensively reviewed in the recent surveys.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
MLBL (Kiros et al., 2014a)	0.01	3.4	4.99	1.02	4.08	62
NeuralTalk (Karpathy & Fei-Fei, 2015)	0.00	0.4	1.34	1.02	3.40	88
NIC (Vinyals et al., 2015)	0.07	10.0	4.51	2.83	10.38	61.5
CRCN (Park & Kim, 2015)	3.49	52.7	8.78	14.29	31.29	16
Story-Flat (Huang et al., 2016)	-	-	7.61	-	-	-
HierarchialRNN (Krause et al., 2017)	-	-	7.72	-	-	-
BARNN (Liu et al., 2017)	-	**54.1**	-	**35.01**	**49.07**	**6**
Adversarial (Wang et al., 2018a)	-	-	**9.90**	-	-	-

Table 103: Language and Retrieval Metrics of different methods on the "Disneyland-Storytelling" (Section 3.3.2) dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
CRCN (Park & Kim, 2015)	-	-	-	9.87	28.74	21
Story-Flat (Huang et al., 2016)	3.50	6.84	10.25	-	-	-
HierarchialRNN (Krause et al., 2017)	3.7	6.51	9.97	-	-	-
Adversarial (Wang et al., 2018a)	**5.16**	**11.35**	**12.32**	-	-	-

Table 104: Language and Retrieval Metrics of different methods on the "SIND" (Section 3.3.3) dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
enc-attn-dec (Xu et al., 2015a)	-	4.96	32.98	-	-	-
h-attn-rank (Yu et al., 2017)	-	7.38	33.94	-	-	-
BARNN (Liu et al., 2017)	-	-	33.32	**24.07**	**44.29**	**9**
AREL-t-100 (Wang et al., 2018a)	14.1	**9.4**	**35.0**	-	-	-

Table 105: Language and Retrieval Metrics of different methods on the "VIST" (Section 3.3.4) dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
seq-seq+context (Gella et al., 2018)	1.20	9.37	33.88	-	-	-

Table 106: Language and Retrieval Metrics of different methods on the "VideoStory" (Section 3.4.1) dataset.

Model	B-4	CIDEr	METEOR	R@1	R@5	MedRank
mRNN (Mao et al., 2014)	11.8	81.3	18.0	5.34	21.23	29
Deep Video-Text (Xu et al., 2015b)	11.5	79.5	17.7	4.72	19.85	31
H-RNN (Yu et al., 2016)	**16.1**	64.6	15.5	-	-	-
ResBRNN (Li et al., 2018)	14.7	94.3	19.6	**7.44**	**25.77**	**22**
ResBRNN-kNN (Li et al., 2018)	15.6	**103.6**	**20.1**	-	-	-

Table 107: Language and Retrieval Metrics of different methods on the "VideoStory-NUS" (Section 3.4.2) dataset.

6.2.1 Image Dialog

To evaluate image dialog, two different datasets presented in the Section 3.7 are used. Retrieval metrics are used as evaluation measures for comparing different methods. Table 108, Table 109 and Table 110 present the benchmark results on each of these datasets separately.

Model	MRR	R@1	R@5	R@10	Mean
LF (Das et al., 2017a)	0.5807	43.82	74.68	84.07	5.78
HRE (Das et al., 2017a)	0.5846	44.67	74.50	84.22	5.72
HREA (Das et al., 2017a)	0.5868	44.82	74.81	84.36	5.66
MN (Das et al., 2017a)	0.5965	45.55	76.22	85.37	5.46
HCIAE-NP-ATT (Lu et al., 2017)	0.6222	48.48	78.75	87.59	4.81
AMEM (Seo et al., 2017)	0.6227	48.53	78.66	87.43	4.86
CoAtt (Wu et al., 2018)	0.6398	50.29	80.71	88.81	4.47
SF (Jain et al., 2018)	0.6242	48.55	78.96	87.75	4.70
SCA (Wu et al., 2018)	0.6398	50.29	80.71	88.81	4.47
CorefNMN (Kottur et al., 2018)	0.641	50.92	80.18	88.81	4.45
GNN (Zheng et al., 2019)	0.6285	48.95	79.65	88.36	4.57
RvA (Niu et al., 2019)	**0.6634**	**52.71**	**82.97**	**90.73**	**3.93**

Table 108: Retrieval Metrics of different **discriminative methods** on "VisDial v0.9" (Section 3.7.1) validation split of the dataset.

Model	MRR	R@1	R@5	R@10	Mean
LF (Das et al., 2017a)	0.5199	41.83	61.78	67.59	17.07
HRE (Das et al., 2017a)	0.5237	42.29	62.18	67.92	17.07
HREA (Das et al., 2017a)	0.5242	42.28	62.33	68.17	16.79
MN (Das et al., 2017a)	0.5259	42.29	62.85	68.88	17.06
HCIAE-NP-ATT (Lu et al., 2017)	0.5386	44.06	63.55	69.24	16.01
CorefNMN (Kottur et al., 2018)	0.535	43.66	63.54	69.93	15.69
CoAtt (Wu et al., 2018)	0.5411	44.32	63.82	69.75	16.47
CoAtt-RL (Wu et al., 2018)	**0.5578**	**46.10**	**65.69**	71.74	14.43
RvA (Niu et al., 2019)	0.5543	45.37	65.27	**72.97**	**10.71**

Table 109: Retrieval Metrics of different **generative methods** on "VisDial v0.9" (Section 3.7.1) validation split of the dataset.

6.2.2 Video Dialog

To evaluate video dialog, the dataset presented in the Section 3.8 is used. Language metrics are used as evaluation measures for comparing different methods. Table 111 present the benchmark results.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF (Das et al., 2017a)	0.5542	40.95	72.45	82.83	5.95	0.4531
LF-att (Das et al., 2017a)	0.5707	42.08	74.83	85.05	5.59	0.4976
HRE (Das et al., 2017a)	0.5416	39.93	70.45	81.50	6.41	0.4546
MN (Das et al., 2017a)	0.5549	40.98	72.30	83.30	5.92	0.4750
MN-att (Das et al., 2017a)	0.5690	42.43	74.00	84.35	5.59	0.4958
CorefNMN (Kottur et al., 2018)	0.615	47.55	78.10	88.80	4.40	0.547
GNN (Zheng et al., 2019)	0.6137	47.33	77.98	87.83	4.57	0.5282
RvA (Niu et al., 2019)	0.6303	49.03	80.40	89.83	4.18	0.5559
Synergistic-ensemble (Guo et al., 2019)	**0.6342**	**49.30**	**80.77**	**90.68**	**3.97**	**0.5788**

Table 110: Retrieval Metrics of different **discriminative methods** on "VisDial v1.0" (Section 3.7.1) test-standard split of the dataset.

Model	B-1	B-2	B-3	B-4	METEOR	CIDEr
Att-base (Hori et al., 2019)	0.273	0.173	0.117	0.084	0.117	0.766
Att-weightshare (Schwartz et al., 2019)	**0.293**	**0.191**	0.133	0.097	0.127	0.923
i3d-rgb-spatial-10 (Schwartz et al., 2019)	0.290	0.190	**0.133**	**0.097**	0.127	0.928
Att-base-beam (Schwartz et al., 2019)	0.285	0.187	0.131	0.096	**0.128**	**0.941**

Table 111: Language metrics of different methods on the "AVSD" (Section 3.8) dataset.

6.3 Visual Reasoning Results

In the following, we present the results achieved with various methods for image and video reasoning separately.

6.3.1 Image Reasoning

To evaluate image reasoning, three different datasets mainly CLEVR, GQA and RAVEN presented in the Section 3.9 is used. Task-specific metrics are used as evaluation measures for comparing different methods. Table 112, Table 113 and Table 114 present the benchmark results on each of these datasets separately.

6.3.2 Video Reasoning

To evaluate video reasoning, COG dataset presented in the Section 3.10 is used. Task-specific metrics are used as evaluation measures for comparing different methods. Table 115 present the benchmark results.

6.4 Visual Referring Expression Results

In the following, we present the results achieved with various methods for image and video referring expression separately.

Model	Count	Exist	CN	QA	CA	Overall
CNN+LSTM+SA+MLP (Johnson et al., 2017a)	59.7	77.9	75.1	80.9	70.8	73.2
N2NMN+700KProgLabel (Hu et al., 2017)	68.5	85.7	84.9	90.0	88.7	83.7
PGEE+700KProgLabel (Johnson et al., 2017b)	92.7	97.1	98.7	98.1	98.9	96.9
CNN+LSTM+RN (Santoro et al., 2017)	90.1	97.8	93.6	97.9	97.1	95.5
ACMN (Cao et al., 2018)	94.2	81.3	81.6	90.5	97.1	89.3
CNN+GRU+FiLM (Perez et al., 2018)	94.3	99.1	96.8	99.1	99.1	97.7
MAC (Hudson & Manning, 2018)	97.2	**99.5**	99.4	99.3	99.5	98.9
TbD+700KProgLabel (Mascharka et al., 2018)	97.6	99.2	99.4	**99.5**	99.6	**99.1**
FinalDestGraph (Haurilet et al., 2019)	91.3	98.6	**99.6**	**99.5**	**99.8**	97.5
LCGN+single-hop (Hu et al., 2019)	-	-	-	-	-	97.9
NS-CL (Mao et al., 2019)	**98.2**	98.8	99.0	99.3	99.1	98.9

Table 112: Comparison of different methods using task-specific metrics (Section 5.2.1) on the "CLEVR" (Section 3.9.1) dataset.

Model	val	test-dev	test
CNN+LSTM (Hudson & Manning, 2019)	49.2	-	46.6
Bottom-up (Anderson et al., 2018)	52.2	-	49.7
MAC (Hudson & Manning, 2018)	57.5	-	54.1
LCGN+single-hop (Hu et al., 2019)	**63.8**	**55.6**	**56.0**

Table 113: Comparison of accuracy (%) scores of different methods on the "GQA" (Section 3.9.3) validation (val), test-dev and test dataset.

Model	Acc	Center	2x2 Grid	3x3 Grid	L-R	U-D	O-IC	O-IG
WReNDRT (Santoro et al., 2018)	15.02	15.38	23.26	29.51	6.99	8.43	8.93	12.35
ResNetDRT (Zhang et al., 2019)	**59.56**	**58.08**	**46.53**	**50.40**	**65.82**	**67.11**	**69.09**	**60.11**
Human (Zhang et al., 2019)	84.41	95.45	81.82	79.55	86.36	81.81	86.36	81.81
PerfectSolver	100	100	100	100	100	100	100	100

Table 114: Comparison of accuracy (%) scores of different methods on the "RAVEN" (Section 3.9.4) dataset.

Model	Atts	Condit	Point	Yes/No	All
WorkMemory (Yang et al., 2018)	-	-	-	-	93.7
QuestionNodes (Haurilet et al., 2019)	73.7	63.5	92.5	57.9	63.3
FinalDestGraph (Haurilet et al., 2019)	**99.2**	**98.4**	**100.0**	**95.0**	**97.2**

Table 115: Comparison of different methods using task-specific metrics (Section 5.2.1) on the "COG" (Section 3.10) dataset.

6.4.1 Image Referring Expression

To evaluate image referring expression, three different datasets mainly RefCOCO, RefCOCO+ and RefCOCOg presented in the Section 3.11 is used. Precision@1 i.e., the precision

	RefCOCO		
Model	val	testA	testB
MMI (Mao et al., 2016)	-	63.15	64.21
Neg. Bag (Nagaraja et al., 2016)	76.90	75.60	78.00
Context (Yu et al., 2016)	76.18	74.39	77.30
CG (Luo & Shakhnarovich, 2017)	-	74.04	73.43
Attributes (Liu et al., 2017)	-	78.85	78.07
CMN (Hu et al., 2017b)	-	75.94	79.57
Reinforcer (Yu et al., 2017a)	79.56	78.95	80.22
VarContext (Zhang et al., 2018b)	-	78.98	82.39
AccumulateAtt (Deng et al., 2018)	81.27	81.17	80.01
ParallelAtt (Zhuang et al., 2018)	81.67	80.81	81.32
MAttNet+ResNet-101 (Yu et al., 2018)	85.65	85.26	84.57
RVG-Tree+ResNet-101 (Hong et al., 2019)	83.48	82.52	82.90
CMRIN+ResNet-101 (Yang et al., 2019)	**86.99**	**87.63**	**84.73**

Table 116: Comparison of Precision@1 (%) scores of different methods for Referring Expression Comprehension on the "RefCOCO" (Section 3.11.1) dataset.

	RefCOCO+			RefCOCOg	
Model	val	testA	testB	val	test
MMI (Mao et al., 2016)	-	48.73	42.13	-	-
Neg Bag (Nagaraja et al., 2016)	-	-	-	-	68.40
Context (Yu et al., 2016)	58.94	61.29	56.24	-	-
CG (Luo & Shakhnarovich, 2017)	-	60.26	55.03	-	-
Attributes (Liu et al., 2017)	-	61.47	57.22	-	-
CMN (Hu et al., 2017b)	-	59.29	59.34	-	-
Reinforcer (Yu et al., 2017a)	62.26	64.60	59.62	71.65	71.92
VariationalContext (Zhang et al., 2018b)	-	62.56	62.90	-	-
AccumulateAttn (Deng et al., 2018)	65.56	68.76	60.63	-	-
ParallelAttn (Zhuang et al., 2018)	64.18	66.31	61.46	-	-
MAttNet+ResNet-101 (Yu et al., 2018)	71.01	75.13	66.17	78.10	78.12
RVG-Tree+ResNet-101 (Hong et al., 2019)	68.86	70.21	65.49	76.82	75.20
CMRIN+ResNet-101 (Yang et al., 2019)	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**

Table 117: Comparison of Precision@1 (%) scores of different methods for Referring Expression Comprehension on the "RefCOCO+" and "RefCOCOg" (Section 3.11.1) dataset.

calculated with Intersection over Union (IoU) ratio between the true and predicted bounding box is used as evaluation measures for comparing different methods. Table 116 and Table 117 present the benchmark results on each of these datasets separately.

6.4.2 Video Referring Expression

To evaluate video referring expression, the dataset presented in the Section 3.12 is used. Top-1 accuracy is used as evaluation measures for comparing different methods. Table 118 present the benchmark results.

Methods	Edgebox	FRCNN	LOP
MNLM (Kiros et al., 2014b)	-	23.954	32.418
VSEM (Liu et al., 2015)	-	24.833	32.961
MCB (Fukui et al., 2016)	-	26.445	33.366
SimModel (Plummer et al., 2017)	4.5	18.431	35.556
WithGaze (Balajee Vasudevan et al., 2018)	-	**47.256**	**47.012**

Table 118: Comparison of Top-1 Accuracy (%) scores of different methods for Referring Expression Comprehension on the "ORGaze" (Section 3.12) dataset. Metrics are Object proposal accuracies referred with Language based Object Proposals (LOP), Faster R-CNN (FRCNN) and EdgeBox (Zitnick & Dollár, 2014).

6.5 Visual Entailment Results

In the following, we present the results achieved with various methods for image entailment.

6.5.1 Image Entailment

To evaluate image entailment, two different datasets mainly SNLI-VE and V-SNLI presented in the Section 3.13 is used. Accuracy is used as evaluation measures for comparing different methods. Table 119, Table 120 and Table 121 present the benchmark results on each of these datasets separately.

Model	Contradiction	Neutral	Entailment	Overall
Relation Network (Santoro et al., 2017)	67.29	68.86	66.50	67.55
Bottom-up (Anderson et al., 2017b)	70.52	**70.96**	65.23	68.90
Top-Down (Anderson et al., 2017b)	69.72	69.33	71.86	70.3
Hypothesis Only (Gururangan et al., 2018)	67.60	67.71	64.83	66.71
Image Captioning (Choi, 2018)	66.25	70.69	66.08	67.67
EVE-ROI (Xie et al., 2019)	67.69	69.45	**74.25**	70.47
EVE-Image (Xie et al., 2019)	**71.56**	70.52	71.39	**71.16**

Table 119: Comparison of accuracy (%) scores of different methods on "SNLI-VE" (Section 3.13) dataset.

6.6 Language-to-Vision Generation Results

In the following, we present the results achieved with various methods for language-to-image and language-to-video generation separately.

Model	Contradiction	Neutral	Entailment	Overall
Hypothesis Only (Bowman et al., 2015)	66.29	66.36	72.65	68.49
LSTM (blind) (Bowman et al., 2015)	79.7	76.79	87.71	81.49
V-LSTM (Anderson et al., 2017b)	71.39	68.06	87.14	75.70
BiMPM (Wang et al., 2017)	86.25	82.79	90.03	86.41
V-BiMPM (Vu et al., 2018)	**87.53**	**82.91**	**90.38**	**86.99**

Table 120: Comparison of accuracy (%) scores of different methods on "V-SNLI" (Section 3.13) dataset.

Model	Contradiction	Neutral	Entailment	Overall
Hypothesis Only (Bowman et al., 2015)	25.29	20.22	31.28	25.57
LSTM (blind) (Bowman et al., 2015)	60.79	50.19	72.12	60.99
V-LSTM (Anderson et al., 2017b)	46.34	32.02	69.09	49.03
BiMPM (Wang et al., 2017)	**77.62**	59.36	80.43	72.55
V-BiMPM (Vu et al., 2018)	76.12	**63.67**	**81.38**	**73.75**

Table 121: Comparison of accuracy (%) scores of different methods on "V-SNLI$_{\text{hard}}$" (Section 3.13) dataset.

6.6.1 Language-to-Image Generation

To evaluate language-to-image generation, three different datasets mainly CUB, Oxford-102 and COCO presented in the Section 3.14 is used. Task-specific evaluation measures are used for comparing different methods. Table 122, Table 123 and Table 124 present the benchmark results on each of these datasets separately.

6.6.2 Language-to-Video Generation

To evaluate language-to-video generation, the TexttoVideo dataset presented in the Section 3.15 is used. Accuracy is used as evaluation measures for comparing different methods. Table 125 present the benchmark results.

6.7 Vision-and-Language Navigation Results

In the following, we present the results achieved with various methods for image-and-language navigation.

6.7.1 Image-and-Language Navigation

To evaluate image-and-language navigation, the dataset presented in the Section 3.16 is used. Task-specific evaluation measures are used for comparing different methods. Table 126, Table 127 and Table 128 present the benchmark results on different splits of the dataset.

6.8 Multimodal Machine Translation Results

In the following, we present the results achieved with various methods for machine translation with an image and a video separately.

Model	Resolution	IS	FID	HR
GAN-INT-CLS (Reed et al., 2016a)	64x64	2.88 ± .04	68.79	2.76 ± .01
	128x128	-	-	-
	256x256	-	-	-
GAWWN (Reed et al., 2016b)	64x64	3.10 ± .03	53.51	-
	128x128	3.62 ± .07	72.65	1.95 ± .02
	256x256	-	-	-
StackGAN (Zhang et al., 2017)	64x64	3.02 ± .03	35.11	-
	128x128	-	-	-
	256x256	3.70 ± .04	51.89	1.29 ± .02
StackGAN++ (Zhang et al., 2018a)	64x64	-	-	-
	128x128	-	-	-
	256x256	4.04 ± .05	15.30	1.19 ± .02
PPGN (Nguyen et al., 2017)	64x64	-	-	-
	128x128	-	-	-
	256x256	-	-	-
AttGAN (Xu et al., 2018)	64x64	-	-	-
	128x128	-	-	-
	256x256	4.36 ± .03	-	-
MirrorGAN (Qiao et al., 2019)	64x64	-	-	-
	128x128	-	-	-
	256x256	4.56 ± .05	-	-

Table 122: Comparison of different methods using generated images of different resolutions with task-specific metrics (Section 5.2.2) on the "CUB" (Section 3.14) dataset. R-precision (%) for 256x256 with AttGAN (53.31) and MirrorGAN (57.67). HR - Human Ranking

6.8.1 Machine Translation with Image

To evaluate machine translation with an image, the dataset presented in the Section 3.17 is used. Language metrics are used as evaluation measures for comparing different methods. Table 129 and Table 130 present the benchmark results on different splits of the dataset.

6.8.2 Machine Translation with Video

To evaluate language-to-video generation, the dataset presented in the Section 3.18 is used. Language metrics are used as evaluation measures for comparing different methods. Table 131 present the benchmark results.

7. Future Directions

The integration of vision and language has come very far since the pioneer methods, especially after the adoption of deep learning techniques. Although the performance of existing methods still needs to catch up with human abilities, the gap is diminishing at a steady rate and there is still ample room for algorithmic improvements. Here, we list several possible future directions that have the potential to advance the overall research area.

Model	Resolution	IS	FID	HR
GAN-INT-CLS (Reed et al., 2016a)	64x64	2.66 ± .03	79.55	1.84 ± .02
	128x128	-	-	-
	256x256	-	-	-
GAWWN (Reed et al., 2016b)	64x64	-	-	-
	128x128	-	-	-
	256x256	-	-	-
StackGAN (Zhang et al., 2017)	64x64	2.73 ± .03	43.02	-
	128x128	-	-	-
	256x256	3.20 ± .01	55.28	1.16 ± .02
StackGAN++ (Zhang et al., 2018a)	64x64	-	-	-
	128x128	-	-	-
	256x256	3.26 ± .01	48.68	1.30 ± .03
PPGN (Nguyen et al., 2017)	64x64	-	-	-
	128x128	-	-	-
	256x256	-	-	-
AttGAN (Xu et al., 2018)	64x64	-	-	-
	128x128	-	-	-
	256x256	-	-	-
MirrorGAN (Qiao et al., 2019)	64x64	-	-	-
	128x128	-	-	-
	256x256	-	-	-

Table 123: Comparison of different methods using generated images of different resolutions with task-specific metrics (Section 5.2.2) on the "Oxford-102" (Section 3.14) dataset.

Leveraging External Knowledge: There is abundant out-of-domain information available which is unpaired with the vision and language task-specific corpora. Leveraging such information available as factual or commonsense knowledge can significantly improve integration of vision and language tasks. There has been prior work shown to assist independent NLP tasks with pretrained language models such as commonsense reasoning (Rajani et al., 2019) and fact predictions (Logan IV et al., 2019). It has also shown prominence for image caption generation (Wu et al., 2017a; Mogadala et al., 2018) and question answering (Shah et al., 2019a; Marino et al., 2019). Extending such ideas to other tasks would be an interesting research direction to explore.

Addressing Large-scale Data Limitations: Most of the approaches designed for tasks that integrate vision and language use large training datasets for training. However, it will become soon harder to design new tasks without having a dataset. To overcome such a situation, approaches need to be designed in such a way where the size of training dataset becomes oblivious. Therefore, trade-off approaches are required where we know how much amount of data is enough to master a certain task. This requires designing methods that leverage neuro-symbolic reasoning systems (Yi et al., 2018; Vedantam et al., 2019) which decide the required amount of the data.

Model	Resolution	IS	FID	HR
GAN-INT-CLS (Reed et al., 2016a)	64x64	7.88 ± .07	60.62	1.82 ± .03
	128x128	-	-	-
	256x256	-	-	-
GAWWN (Reed et al., 2016b)	64x64	-	-	-
	128x128	-	-	-
	256x256	-	-	-
StackGAN (Zhang et al., 2017)	64x64	8.35 ± .11	33.88	-
	128x128	-	-	-
	256x256	8.45 ± .03	74.05	1.18 ± .03
StackGAN++ (Zhang et al., 2018a)	64x64	-	-	-
	128x128	-	-	-
	256x256	8.30 ± .10	81.59	1.55 ± .05
PPGN (Nguyen et al., 2017)	64x64	-	-	-
	128x128	-	-	-
	256x256	9.58 ± .21	-	-
AttGAN (Xu et al., 2018)	64x64	-	-	-
	128x128	-	-	-
	256x256	25.89 ± .47	-	-
MirrorGAN (Qiao et al., 2019)	64x64	-	-	-
	128x128	-	-	-
	256x256	26.47 ± .41	-	-

Table 124: Comparison of different methods using generated images of different resolutions with task-specific metrics (Section 5.2.2) on the "COCO" (Section 3.14) dataset. R-precision (%) for 256x256 with AttGAN (72.13) and MirrorGAN (74.52).

Model	Accuracy
DT2V-baseline (Li et al., 2018b)	0.101
PT2V (Reed et al., 2016a)	0.134
GT2V (Li et al., 2018b)	0.192
T2V (Li et al., 2018b)	**0.426**

Table 125: Comparison of accuracy (%) scores of different methods on "TexttoVideo" (Section 3.15) dataset.

Combining Multiple Tasks There are tasks which can share ideas from each other. For example, visual referring expression comprehension can be seen as a visual dialog task (de Vries et al., 2017) where a sequence of questions is used to refer to an object in an image. Similarly, image caption generation can be seen as the visual referring expression generation task (Mao et al., 2016).

Novel Neural Architectures for Representation: Previously, the de facto neural network architectures used for language and vision representation were RNNs and CNNs respectively. However, with the introduction of novel ideas which address the limitations of them

Model	PL	NE	OSR	SR	SPL
Random	9.89	9.79	18.3	13.2	12
Seq-to-Seq (Anderson et al., 2018)	8.13	7.85	26.6	20.4	18
RPA (Wang et al., 2018)	9.15	7.53	32.5	25.3	23
Speaker-Follower (Fried et al., 2018)	14.82	6.62	44.0	35.0	28
Self-Monitoring (Ma et al., 2019a)	18.0	-	-	48.0	35
RCM (Wang et al., 2018c)	15.22	6.01	50.8	43.1	35
BackTranslation-Single (Tan et al., 2019)	11.7	-	-	**51.5**	**47**
TacticalRewind-Greedy (Ke et al., 2019)	22.08	5.14	-	54	41
BackTranslation-PreExplore (Tan et al., 2019)	9.79	-	-	63.9	61
BackTranslation-Beam (Tan et al., 2019)	687	-	-	68.9	1
TacticalRewind-Beam (Ke et al., 2019)	196.53	4.29	-	61.0	3

Table 126: Comparison of different methods using task-specific metrics (Section 5.2.3) on "R2R" (Section 3.16.1) test dataset.

Model	PL	NE	OSR	SR	SPL
Speaker-Follower (Fried et al., 2018)	-	3.36	73.8	66.4	-
RCM+SIL (Wang et al., 2018c)	10.13	2.78	79.7	73.0	-
BackTranslation-Single (Tan et al., 2019)	11.0	3.99	-	62.1	59
TacticalRewind-Greedy (Ke et al., 2019)	-	-	-	-	-
BackTranslation-PreExplore (Tan et al., 2019)	9.92	4.84	-	54.7	52
BackTranslation-Beam (Tan et al., 2019)	703	2.52	-	75.7	1
TacticalRewind-Beam (Ke et al., 2019)	188.6	3.13	-	70.0	4

Table 127: Comparison of different methods using task-specific metrics (Section 5.2.3) on "R2R" (Section 3.16.1) seen validation dataset.

Model	PL	NE	OSR	SR	SPL
Speaker-Follower (Fried et al., 2018)	-	3.36	73.8	66.4	-
RCM+SIL (Wang et al., 2018c)	10.13	2.78	79.7	73.0	-
BackTranslation-Single (Tan et al., 2019)	10.7	5.22	-	52.2	48
TacticalRewind-Greedy (Ke et al., 2019)	21.17	4.97	-	56.0	43
BackTranslation-PreExplore (Tan et al., 2019)	9.57	3.78	-	64.5	61
BackTranslation-Beam (Tan et al., 2019)	663	3.08	-	69.0	1
TacticalRewind-Beam (Ke et al., 2019)	224.42	4.03	-	63.0	2

Table 128: Comparison of different methods using task-specific metrics (Section 5.2.3) on "R2R" (Section 3.16.1) unseen validation dataset.

either theoretically or computationally there is interest to adopt these new techniques. For example, the Transformer (Vaswani et al., 2017) architecture which is used extensively new for pure NLP tasks, may see its high adoption for integration of language and vision tasks. It has already shown its applicability for the image caption generation (Sharma et al., 2018). Similarly, graph neural networks (Scarselli et al., 2008; Kipf & Welling, 2016; Battaglia et al., 2018) introduced to tackle graph-structured data has already shown its prominence

Results of Different Methods				
Model	Language	en → de	en → fr	en → cs
DoubleAtt (Calixto et al., 2017)	BLEU	36.5	-	-
	METEOR	55.0	-	-
GVF (Calixto & Liu, 2017)	BLEU	37.3	-	-
	METEOR	55.1	-	-
Imagination (Elliott & Kádár, 2017)	BLEU	36.8	-	-
	METEOR	55.8	-	-
Lium-cvc-ensemble (Caglayan et al., 2017)	BLEU	41.0	56.7	-
	METEOR	60.5	73.0	-
$VMMT_F$ (Calixto et al., 2018)	BLEU	37.6	-	-
	METEOR	56.0	-	-
CUNI-ensemble (Helcl et al., 2018)	BLEU	42.6	62.8	35.9
	METEOR	59.4	77.0	32.7

Table 129: Machine Translation with image on "Multi30K" [2016 (en → de), 2017 (en → fr), 2018 (en → cs)] test set.

Results of Different Methods				
Model	Language	en → de	en → fr	en → cs
CUNI-single (Helcl et al., 2018)	BLEU	32.5	40.6	31.8
	METEOR	52.3	61.0	30.6
MeMAD (Grönroos et al., 2018)	BLEU	38.5	44.1	-
	METEOR	56.6	64.3	-

Table 130: Machine Translation with image on "Multi30K" [2018 (en → de, en → fr, en → cs)] test set.

Model	B-4	METEOR
NMT+LSTM VI (Wang et al., 2019) [English → Chinese]	30.20	-
NMT+LSTM VI (Wang et al., 2019) [Chinese → English]	27.18	-

Table 131: Comparison of different methods on "VATEX" dataset.

in visual reasoning (Haurilet et al., 2019). Extending such ideas to other tasks would be interesting direction to explore.

Image vs Video: Most of the integration of vision and language research concentrated mostly on images. This is clearly visible from the datasets and methods availability for image and language integration. Nevertheless, although a complex task similar attention needs to be embraced for videos. There is only one dataset available for tasks such as Video Dialog (Section 2.4.2), Video Reasoning (Section 2.5.2), Video Referring Expression (Section 2.6.2), Language-to-Video Generation (Section 2.8.2) and Machine Translation with Videos (Section 2.10.2). While, tasks such as Visual Entailment (Section 2.7) and Vision-and-Language Navigation (Section 2.9) even lack video datasets.

Automatic Evaluation Measures: There are automatic evaluation measures designed for several vision and language tasks. However, most of them are adaptations from standalone NLP tasks such as machine translation. For example, BLEU and METEOR used for visual caption generation and storytelling have shown not to correlate well with human judgements (Bernardi et al., 2016). SPICE designed specifically for visual caption generation is dependent on parsing is not adaptable for other tasks such as storytelling. This shows us a promising research direction to develop measures usable for several tasks. Similarly, language-to-vision generation although having quantitative measures is typically dependent on human evaluation. It needs to adopt novel techniques for effective quantitative evaluation. Other tasks such as vision-and-language navigation, visual reasoning have specific measures for evaluation which can be improved further.

8. Conclusion

In this survey, we discussed recent trends in integration of vision and language research. We reviewed and analyzed ten different prominent tasks by presenting novel methods used for their research. In particular, we analyzed how the tasks are designed bottom-up and found what is the commonality between them interms of language and vision representation. In addition, we provided a brief review of the datasets, evaluation measures, relative performance achieved with state of the art methods and finally concluded with some future directions for integration of language and vision research.

Compared to the stand alone research done in the independent fields of CV and NLP, combining both of them with advanced machine learning techniques could lead to more intelligent and sustainable systems. Making them easily accessible can also have direct commercial and societal impact. However, despite significant progress observed in many tasks, evaluation of systems show that they still fall below human performance. This shows that there is large scope for improvement. To be specific, designing novel evaluation measures and architectures which can deal with the complexity of vision and language integration problem adequately can address the challenge.

We believe that our survey will help to systematize future research papers and also explore the unresolved problems present in the integration of vision and language research.

Acknowledgments

This work was supported by the German Research Foundation (DFG) as part of SFB1102. Thanks to Marius Mosbach for his insightful comments on the draft.

References

Aafaq, N., Gilani, S. Z., Liu, W., & Mian, A. (2018). Video description: a survey of methods, datasets and evaluation metrics. *arXiv preprint arXiv:1806.00186*.

Aditya, S., Saha, R., Yang, Y., & Baral, C. (2019). Spatial knowledge distillation to aid visual reasoning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 227–235. IEEE.

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pp. 4971–4980. IEEE Computer Society.

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., & Batra, D. (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, *123*(1), 4–31.

Agrawal, H., Chandrasekaran, A., Batra, D., Parikh, D., & Bansal, M. (2016). Sort story: Sorting jumbled images and captions into stories. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 925–931.

Alamri, H., Cartillier, V., Das, A., Wang, J., Lee, S., Anderson, P., Essa, I., Parikh, D., Batra, D., Cherian, A., et al. (2019a). Audio-visual scene-aware dialog. *arXiv preprint arXiv:1901.09107*.

Alamri, H., Hori, C., Marks, T. K., Batr, D., & Parikh, D. (2019b). Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 workshop at AAAI*.

Anderson, P., Chang, A., Chaplot, D. S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al. (2018). On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398. Springer.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2017a). Guided open vocabulary image captioning with constrained beam search. In *EMNLP*.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2017b). Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., & van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016a). Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016b). Neural module networks. In *CVPR*, pp. 39–48. IEEE Computer Society.

Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433.

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

Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304.

Balajee Vasudevan, A., Dai, D., & Van Gool, L. (2018). Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4129–4138.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., & Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Baumann, A., Boltz, M., Ebling, J., Koenig, M., Loos, H., Merkel, M., Niem, W., Warzelhan, J., & Yu, J. (2008). A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, *2008*(1), 824726.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures.. *J. Artif. Intell. Res.(JAIR)*, *55*, 409–442.

Blösch, M., Weiss, S., Scaramuzza, D., & Siegwart, R. (2010). Vision based mav navigation in unknown and unstructured environments. In *2010 IEEE International Conference on Robotics and Automation*, pp. 21–28. IEEE.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2).

Burke, H. R. (1958). Raven's progressive matrices: A review and critical evaluation. *The Journal of Genetic Psychology*, *93*(2), 199–228.

Cadene, R., Ben-Younes, H., Cord, M., & Thome, N. (2019). Murel: Multimodal relational reasoning for visual question answering. *arXiv preprint arXiv:1902.09487*.

Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., & Van de Weijer, J. (2017). Lium-cvc submissions for wmt17 multimodal translation task. *arXiv preprint arXiv:1707.04481*.

Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.

Calixto, I., & Liu, Q. (2017). Incorporating global visual features into attention-based neural machine translation.. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 992–1003.

Calixto, I., Liu, Q., & Campbell, N. (2017). Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1913–1924.

Calixto, I., Rios, M., & Aziz, W. (2018). Latent visual cues for neural machine translation. *arXiv preprint arXiv:1811.00357*.

Cao, Q., Liang, X., Li, B., Li, G., & Lin, L. (2018). Visual question reasoning on general dependency tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7249–7257.

Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P. S. (2016). Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1445–1454. ACM.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308.

Chang, S., Yang, J., Park, S., & Kwak, N. (2018). Broadcasting convolutional network for visual relational reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 754–769.

Chen, D. L., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 190–200. Association for Computational Linguistics.

Chen, H., Suhr, A., Misra, D., Snavely, N., & Artzi, Y. (2019). Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12538–12547.

Chen, T.-H., Liao, Y.-H., Chuang, C.-Y., Hsu, W.-T., Fu, J., & Sun, M. (2017). Show, adapt and tell: Adversarial training of cross-domain image captioner. *arXiv preprint arXiv:1705.00930*.

Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2422–2431.

Cheng, Y., Gan, Z., Li, Y., Liu, J., & Gao, J. (2018). Sequential attention gan for interactive image editing via dialogue. *arXiv preprint arXiv:1812.08352*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choi, Y. (2018). Image captioning pytorch implementation..

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 613–622.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cirik, V., Berg-Kirkpatrick, T., & Morency, L.-P. (2018). Using syntax to ground referring expressions in natural images. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., & Bobrow, D. G. (2003). Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning*.

Cornia, M., Baraldi, L., & Cucchiara, R. (2018). Show, control and tell: A framework for generating controllable and grounded captions. *arXiv preprint arXiv:1811.10652*.

Dai, B., & Lin, D. (2017). Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pp. 898–907.

Dai, B., Lin, D., Urtasun, R., & Fidler, S. (2017). Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*.

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018a). Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2054–2063.

Das, A., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018b). Neural modular control for embodied question answering. In *CoRL*, Vol. 87 of *Proceedings of Machine Learning Research*, pp. 53–62. PMLR.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., & Batra, D. (2017a). Visual dialog. In *CVPR*, pp. 1080–1089. IEEE Computer Society.

Das, A., Kottur, S., Moura, J. M. F., Lee, S., & Batra, D. (2017b). Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, pp. 2970–2979. IEEE Computer Society.

Das, P., Xu, C., Doell, R. F., & Corso, J. J. (2013). A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, pp. 2634–2641. IEEE Computer Society.

Dash, A., Gamboa, J. C. B., Ahmed, S., Liwicki, M., & Afzal, M. Z. (2017). Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*.

De Mulder, W., Bethard, S., & Moens, M.-F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, *30*(1), 61–98.

de Vries, H., Shuster, K., Batra, D., Parikh, D., Weston, J., & Kiela, D. (2018). Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, pp. 4466–4475. IEEE Computer Society.

Delbrouck, J.-B., & Dupont, S. (2017a). An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 910–919.

Delbrouck, J.-B., & Dupont, S. (2017b). Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084*.

Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., & Tan, M. (2018). Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7746–7755.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee.

Deshpande, A., Aneja, J., Schwing, A., & Forsyth, D. A. (2018). Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805. 12589*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., Szlam, A., & Weston, J. (2015). Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634.

El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L. E., Kahou, S. E., Bengio, Y., & Taylor, G. W. (2018). Keep drawing it: Iterative language-based image generation and editing. *arXiv preprint arXiv:1811.09845*.

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978.

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177*.

Elliott, D., Frank, S., & Hasler, E. (2015). Multi-language image description with neural sequence models. *CoRR, abs/1510.04709*.

Elliott, D., & Kádár, À. (2017). Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, pp. 130–141.

Elliott, D., & Keller, F. (2013). Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., et al. (2015). From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1473–1482.

Farhadi, A., Hejrati, S. M. M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. A. (2010). Every picture tells a story: Generating sentences from images. In *ECCV (4)*, Vol. 6314 of *Lecture Notes in Computer Science*, pp. 15–29. Springer.

Ferraro, F., Mostafazadeh, N., Vanderwende, L., Devlin, J., Galley, M., Mitchell, M., et al. (2015). A survey of current datasets for vision and language research. *arXiv preprint arXiv:1506.06833*.

FitzGerald, N., Artzi, Y., & Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1914–1925.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., & Darrell, T. (2018). Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pp. 3318–3329.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gan, C., Gan, Z., He, X., Gao, J., & Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3137–3146.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pp. 2296–2304.

Gao, L., Chen, D., Song, J., Xu, X., Zhang, D., & Shen, H. T. (2019). Perceptual pyramid adversarial networks for text-to-image synthesis..

Gao, L., Guo, Z., Zhang, H., Xu, X., & Shen, H. T. (2017). Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, *19*(9), 2045–2055.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423.

Gella, S., & Keller, F. (2017). An analysis of action recognition datasets for language and vision tasks. *arXiv preprint arXiv:1704.07129*.

Gella, S., Lewis, M., & Rohrbach, M. (2018). A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 968–974.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, *112*(12), 3618–3623.

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 410–419. Association for Computational Linguistics.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680.

Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2018). Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4089–4098.

Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2019). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, *127*(4), 398–414.

Graham, Y., Awad, G., & Smeaton, A. (2018). Evaluation of automatic video captioning using direct assessment. *PloS one*, *13*(9), e0202789.

Grönroos, S.-A., Huet, B., Kurimo, M., Laaksonen, J., Merialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., et al. (2018). The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.

Gu, J., Cai, J., Wang, G., & Chen, T. (2018). Stack-captioning: Coarse-to-fine learning for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gu, J., Wang, G., Cai, J., & Chen, T. (2017). An empirical study of language cnn for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1222–1231.

Guo, D., Xu, C., & Tao, D. (2019). Image-question-answer synergistic network for visual dialog. *arXiv preprint arXiv:1902.09774*.

Guo, G., Zhai, S., Yuan, F., Liu, Y., & Wang, X. (2018). Vse-ens: Visual-semantic embeddings with efficient negative sampling. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Guo, L., Liu, J., Yao, P., Li, J., & Lu, H. (2019). Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4204–4213.

Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., & Kembhavi, A. (2018). Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 598–613.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Harabagiu, S. M., Pasca, M. A., & Maiorano, S. J. (2000). Experiments with open-domain textual question answering. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Haurilet, M., Roitberg, A., & Stiefelhagen, R. (2019). It is not about the journey; it is about the destination: Following soft paths under question-guidance for visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Helcl, J., Libovickỳ, J., & Variš, D. (2018). Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., & Darrell, T. (2016). Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, pp. 1–10.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637.

Hinz, T., Heinrich, S., & Wermter, S. (2019). Generating multiple objects at spatially distinct locations. *arXiv preprint arXiv:1901.00686*.

Hitschler, J., Schamoni, S., & Riezler, S. (2016). Multimodal pivots for image caption translation. *arXiv preprint arXiv:1601.03916*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899.

Hong, R., Liu, D., Mo, X., He, X., & Zhang, H. (2019). Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*.

Hong, S., Yang, D., Choi, J., & Lee, H. (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7986–7994.

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T. K., Cartillier, V., Lopes, R. G., Das, A., et al. (2019). End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2352–2356. IEEE.

Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., & Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pp. 4193–4202.

Hossain, M., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, *51*(6), 118.

Hsu, C.-C., Chen, S.-M., Hsieh, M.-H., & Ku, L.-W. (2018). Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. *arXiv preprint arXiv:1805.11867*.

Hu, R., Andreas, J., Darrell, T., & Saenko, K. (2018). Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 53–69.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., & Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 804–813.

Hu, R., Rohrbach, A., Darrell, T., & Saenko, K. (2019). Language-conditioned graph networks for relational reasoning. *arXiv preprint arXiv:1905.04405*.

Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2017a). Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1115–1124.

Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2017b). Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1115–1124.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., & Darrell, T. (2016). Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555–4564.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., & Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Vol. 2, pp. 639–645.

Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., & He, X. (2018). Hierarchically structured reinforcement learning for topically coherent visual story generation. *arXiv preprint arXiv:1805.08191*.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al. (2016). Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Hudson, D. A., & Manning, C. D. (2019). GQA: a new dataset for compositional question answering over real-world images. *CoRR*, *abs/1902.09506*.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

Jabri, A., Joulin, A., & Van Der Maaten, L. (2016). Revisiting visual question answering baselines. In *European conference on computer vision*, pp. 727–739. Springer.

Jain, U., Lazebnik, S., & Schwing, A. G. (2018). Two can play this game: Visual dialog with discriminative question generation and answering. In *CVPR*, pp. 5754–5763. IEEE Computer Society.

Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2758–2766.

Järvelin, K., & Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41–48. ACM.

Jia, X., Gavves, E., Fernando, B., & Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2407–2415.

Jin, J., Fu, K., Cui, R., Sha, F., & Zhang, C. (2015). Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*.

Jin, Q., Chen, J., Chen, S., Xiong, Y., & Hauptmann, A. (2016). Describing videos using multi-modal fusion. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1087–1091. ACM.

Jing, L., & Tian, Y. (2019). Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint arXiv:1902.06162*.

Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017a). Clevr: A diagnostic dataset for compositional language and elementary

visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017b). Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2989–2998.

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

Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3–20.

Kafle, K., Shrestha, R., & Kanan, C. (2019). Challenges and prospects in vision and language research. *arXiv preprint arXiv:1904.09317*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798.

Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J., Choi, Y., & Srinivasa, S. (2019). Tactical rewind: Self-correction via backtracking in vision-and-language navigation. *arXiv preprint arXiv:1903.02547*.

Khoreva, A., Rohrbach, A., & Schiele, B. (2018). Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pp. 123–141. Springer.

Kim, D.-J., Choi, J., Oh, T.-H., & Kweon, I. S. (2019). Dense relational captioning: Triple-stream networks for relationship-based captioning. *arXiv preprint arXiv:1903.05942*.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014a). Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 595–603.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kottur, S., Moura, J. M. F., Parikh, D., Batra, D., & Rohrbach, M. (2018). Visual coreference resolution in visual dialog using neural module networks. In *ECCV (15)*, Vol. 11219 of *Lecture Notes in Computer Science*, pp. 160–178. Springer.

Kottur, S., Moura, J. M., Parikh, D., Batra, D., & Rohrbach, M. (2019). Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 582–595.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 317–325.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Niebles, J. (2017a). Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017b). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., & Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Lei, J., Yu, L., Bansal, M., & Berg, T. (2018). Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379.

Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2019). Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*.

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation..

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2018). Video storytelling. *arXiv preprint arXiv:1807.09418*.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 220–228. Association for Computational Linguistics.

Li, X., Zhou, Z., Chen, L., & Gao, L. (2018). Residual attention-based lstm for video captioning. *World Wide Web*, 1–16.

Li, Y., Yao, T., Pan, Y., Chao, H., & Mei, T. (2019). Pointing novel objects in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12497–12506.

Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., & Gao, J. (2018a). Storygan: A sequential conditional gan for story visualization. *arXiv preprint arXiv:1812.02784*.

Li, Y., Min, M. R., Shen, D., Carlson, D., & Carin, L. (2018b). Video generation from text. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liang, X., Hu, Z., Zhang, H., Gan, C., & Xing, E. P. (2017). Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3362–3371.

Libovickỳ, J., & Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, pp. 196–202.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer.

Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., & Yuille, A. (2017). Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1271–1280.

Liu, D., Bober, M., & Kittler, J. (2019). Visual semantic information pursuit: A survey. *arXiv preprint arXiv:1903.05434*.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2018). Toward abstractive summarization using semantic representations. *arXiv preprint arXiv:1805.10399*.

Liu, J., Wang, L., & Yang, M.-H. (2017). Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4856–4864.

Liu, R., Liu, C., Bai, Y., & Yuille, A. (2019). Clevr-ref+: Diagnosing visual reasoning with referring expressions. *arXiv preprint arXiv:1901.00850*.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pp. 873–881.

Liu, W., Mei, T., Zhang, Y., Che, C., & Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3707–3715.

Liu, Y., Fu, J., Mei, T., & Chen, C. W. (2017). Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Logan IV, R. L., Liu, N. F., Peters, M. E., Gardner, M., & Singh, S. (2019). BarackâĂŹs wife hillary: Using knowledge graphs for fact-aware language modeling..

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Long, X., Gan, C., & de Melo, G. (2018). Video captioning with multi-faceted attention. *Transactions of the Association of Computational Linguistics*, *6*, 173–184.

Lu, J., Kannan, A., Yang, J., Parikh, D., & Batra, D. (2017). Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, pp. 313–323.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2016). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2018). Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228.

Luo, R., & Shakhnarovich, G. (2017). Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7102–7111.

Ma, C.-Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., & Xiong, C. (2019a). Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*.

Ma, C.-Y., Wu, Z., AlRegib, G., Xiong, C., & Kira, Z. (2019b). The regretful agent: Heuristic-aided navigation through progress estimation. *arXiv preprint arXiv:1903.01602*.

Ma, Y.-F., Lu, L., Zhang, H.-J., & Li, M. (2002). A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pp. 533–542. ACM.

MacMahon, M., Stankiewicz, B., & Kuipers, B. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, *2*(6), 4.

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pp. 1682–1690.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pp. 1–9.

Manjunatha, V., Saini, N., & Davis, L. S. (2018). Explicit bias discovery in visual question answering models. *arXiv preprint arXiv:1811.07789*.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100–103.

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015). Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision..

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20.

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

Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204.

Mascharka, D., Tran, P., Soklaski, R., & Majumdar, A. (2018). Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4942–4950.

Mathews, A., Xie, L., & He, X. (2018). Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8591–8600.

Mathews, A. P., Xie, L., & He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mazaheri, A., Zhang, D., & Shah, M. (2017). Video fill in the blank using lstms with spatial-temporal attentions. In *ICCV*, pp. 1416–1425. IEEE Computer Society.

Messina, N., Amato, G., Carrara, F., Falchi, F., & Gennaro, C. (2018). Learning relationship-aware visual features. In *European Conference on Computer Vision*, pp. 486–501. Springer.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., & Artzi, Y. (2018a). Mapping instructions to actions in 3d environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2667–2678.

Misra, I., Girshick, R., Fergus, R., Hebert, M., Gupta, A., & van der Maaten, L. (2018b). Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., & Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756. Association for Computational Linguistics.

Mitchell, M., Van Deemter, K., & Reiter, E. (2013). Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Miyazaki, T., & Shimizu, N. (2016). Cross-lingual image caption generation.. In *ACL (1)*.

Moens, M.-F., Specia, L., & Tuytelaars, T. (2019). Joint processing of language and visual data for better automated understanding (dagstuhl seminar 19021).. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Mogadala, A. (2015). Polylingual multimodal learning. In *ECML PKDD Doctoral Consortium*, p. 155. Citeseer.

Mogadala, A., Bista, U., Xie, L., & Rettinger, A. (2018). Knowledge guided attention and inference for describing images containing unseen objects. In *European Semantic Web Conference*, pp. 415–429. Springer.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G. P., & Vanderwende, L. (2017). Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP(1)*, pp. 462–472. Asian Federation of Natural Language Processing.

Motwani, T. S., & Mooney, R. J. (2012). Improving video activity recognition using object recognition and text mining.. In *ECAI*, Vol. 1, p. 2.

Nagaraja, V. K., Morariu, V. I., & Davis, L. S. (2016). Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pp. 792–807. Springer.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Nam, S., Kim, Y., & Kim, S. J. (2018). Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pp. 42–51.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., & Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4467–4477.

Nguyen, K., Dey, D., Brockett, C., & Dolan, B. (2019). Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12527–12537.

Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., & Wen, J.-R. (2019). Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6679–6688.

Pan, Y., Yao, T., Li, H., & Mei, T. (2017). Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6504–6512.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics.

Park, C. C., & Kim, G. (2015). Expressing an image stream with a sequence of natural sentences. In *Advances in neural information processing systems*, pp. 73–81.

Pasunuru, R., & Bansal, M. (2017a). Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 1273–1283.

Pasunuru, R., & Bansal, M. (2017b). Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 979–985.

Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2016). Areas of attention for image captioning. *arXiv preprint arXiv:1612.01033*.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., & Lazebnik, S. (2017). Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1928–1937.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649.

Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. *arXiv preprint arXiv:1903.05854*.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Rajani, N. F., McCann, B., Xiong, C., & Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Ramanishka, V., Das, A., Park, D. H., Venugopalan, S., Hendricks, L. A., Rohrbach, M., & Saenko, K. (2016). Multimodal video description. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1092–1096. ACM.

Redmon, J., & Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016a). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016b). Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pp. 217–225.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, *1*, 25–36.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge university press.

Ren, M., Kiros, R., & Zemel, R. (2015a). Exploring models and data for image question answering. In *Advances in neural information processing systems*, pp. 2953–2961.

Ren, S., He, K., Girshick, R., & Sun, J. (2015b). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99.

Ren, S., He, K., Girshick, R., & Sun, J. (2015c). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2016). Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*.

Rickert, M., Foster, M. E., Giuliani, M., By, T., Panin, G., & Knoll, A. (2007). Integrating language, vision and action for human robot dialog systems. In *International Conference on Universal Access in Human-Computer Interaction*, pp. 987–995. Springer.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016a). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pp. 817–834. Springer.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2016b). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pp. 817–834. Springer.

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., & Schiele, B. (2014). Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pp. 184–195. Springer.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In *CVPR*, pp. 3202–3212. IEEE Computer Society.

Rohrbach, M., Amin, S., Andriluka, M., & Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1194–1201. IEEE.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013). Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 433–440.

Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., & Schiele, B. (2012). Script data for attribute-based recognition of composite activities. In *European Conference on Computer Vision*, pp. 144–157. Springer.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., & Metze, F. (2018). How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Santoro, A., Hill, F., Barrett, D., Morcos, A., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pp. 4477–4486.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pp. 4967–4976.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

Schamoni, S., Hitschler, J., & Riezler, S. (2018). A dataset and reranking method for multimodal mt of user-generated image captions. *Vol. 1: MT ResearchersâĂŹ Track*, 140.

Schwartz, I., Schwing, A., & Hazan, T. (2019). A simple baseline for audio-visual scene-aware dialog. *arXiv preprint arXiv:1904.05876*.

Seo, P. H., Lehrmann, A. M., Han, B., & Sigal, L. (2017). Visual reference resolution using attention memory for visual dialog. In *NIPS*, pp. 3722–3732.

Shah, M., Chen, X., Rohrbach, M., & Parikh, D. (2019a). Cycle-consistency for robust visual question answering. *CoRR*, *abs/1902.05660*.

Shah, S., Mishra, A., Yadati, N., & Talukdar, P. P. (2019b). Kvqa: Knowledge-aware visual question answering.. AAAI.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565.

Shetty, R., Rohrbach, M., Hendricks, L. A., Fritz, M., & Schiele, B. (2017). Speaking the same language: Matching machine to human captions by adversarial training. *arXiv preprint arXiv:1703.10476*.

Shetty, R., Tavakoli, H. R., & Laaksonen, J. (2018). Image and video captioning with augmented neural architectures. *IEEE MultiMedia*, *25*(2), 34–46.

Shi, J., Zhang, H., & Li, J. (2019). Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shuster, K., Humeau, S., Hu, H., Bordes, A., & Weston, J. (2018). Engaging image captioning via personality. *arXiv preprint arXiv:1810.10665*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., & Rohrbach, M. (2019). Towards vqa models that can read. In *CVPR*. IEEE Computer Society.

Sinopoli, B., Micheli, M., Donato, G., & Koo, T.-J. (2001). Vision based navigation for an unmanned aerial vehicle. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, Vol. 2, pp. 1757–1764. IEEE.

Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., & Shen, H. T. (2017). Hierarchical lstm with adjusted temporal attention for video captioning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2737–2743. AAAI Press.

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description.. In *WMT*, pp. 543–553.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pp. 843–852.

Storks, S., Gao, Q., & Chai, J. Y. (2019). Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Strub, F., de Vries, H., Mary, J., Piot, B., Courville, A. C., & Pietquin, O. (2017). End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*, pp. 2765–2771. ijcai.org.

Strub, F., Seurin, M., Perez, E., De Vries, H., Mary, J., Preux, P., & CourvilleO-livier Pietquin, A. (2018). Visual reasoning with multi-hop feature modulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–800.

Strzalkowski, T., & Harabagiu, S. (2006). *Advances in open domain question answering*, Vol. 32. Springer Science & Business Media.

Su, Y., Fan, K., Bach, N., Kuo, C.-C. J., & Huang, F. (2018). Unsupervised multi-modal neural machine translation. *arXiv preprint arXiv:1811.11365*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Tan, H., Yu, L., & Bansal, M. (2019). Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621.

Tapaswi, M., Bäuml, M., & Stiefelhagen, R. (2015). Book2movie: Aligning video scenes with book chapters. In *CVPR*, pp. 1827–1835. IEEE Computer Society.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas, J. A. (2014). *Meaning in interaction: An introduction to pragmatics*. Routledge.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., & Mooney, R. (2014). Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1218–1227.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26–31.

Torabi, A., Pal, C., Larochelle, H., & Courville, A. (2015). Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., & Paluri, M. (2014). C3d: generic features for video analysis. *CoRR, abs/1412.0767*, *2*(7), 8.

Tsai, Y.-H. H., Huang, L.-K., & Salakhutdinov, R. (2017). Learning robust visual-semantic embeddings. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3591–3600. IEEE.

Tu, K., Meng, M., Lee, M. W., Choe, T. E., & Zhu, S. C. (2014). Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, *21*(2), 42–70.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Vedantam, R., Desai, K., Lee, S., Rohrbach, M., Batra, D., & Parikh, D. (2019). Probabilistic neural-symbolic models for interpretable visual question answering. *CoRR, abs/1902.07864*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575.

Venugopalan, S., Hendricks, L. A., Mooney, R., & Saenko, K. (2016). Improving lstm-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729*.

Venugopalan, S., Hendricks, L. A., Rohrbach, M., Mooney, R., Darrell, T., & Saenko, K. (2017). Captioning images with diverse objects. In *CVPR*.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164.

Vogel, A., & Jurafsky, D. (2010). Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 806–814. Association for Computational Linguistics.

Vu, H., Greco, C., Erofeeva, A., Jafaritazehjan, S., Linders, G., Tanti, M., Testoni, A., Bernardi, R., & Gatt, A. (2018). Grounded textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2354–2368.

Wang, B., Ma, L., Zhang, W., & Liu, W. (2018). Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7622–7631.

Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016). Image captioning with deep bidirectional lstms. In *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 988–997. ACM.

Wang, J., Fu, J., Tang, J., Li, Z., & Mei, T. (2018a). Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wang, J., Madhyastha, P. S., & Specia, L. (2018b). Object counts! bringing explicit detections back into image captioning. In *NAACL-HLT*, pp. 2180–2193. Association for Computational Linguistics.

Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2019). Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 394–407.

Wang, X., Chen, W., Wang, Y.-F., & Wang, W. Y. (2018a). No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 899–909.

Wang, X., Chen, W., Wu, J., Wang, Y.-F., & Yang Wang, W. (2018b). Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4213–4222.

Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.-F., Wang, W. Y., & Zhang, L. (2018c). Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. *arXiv preprint arXiv:1811.10092*.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., & Wang, W. Y. (2019). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *arXiv preprint arXiv:1904.03493*.

Wang, X., Xiong, W., Wang, H., & Yang Wang, W. (2018). Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 37–53.

Wang, Z., Hamza, W., & Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Whitehead, S., Ji, H., Bansal, M., Chang, S.-F., & Voss, C. (2018). Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3992–4001.

Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017a). Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2017b). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*, 21–40.

Wu, Q., Wang, P., Shen, C., Reid, I. D., & van den Hengel, A. (2018). Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, pp. 6106–6115. IEEE Computer Society.

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Xu, H., Li, B., Ramanishka, V., Sigal, L., & Saenko, K. (2019). Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 396–405. IEEE.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057.

Xu, R., Xiong, C., Chen, W., & Corso, J. J. (2015b). Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324.

Yang, G. R., Ganichev, I., Wang, X.-J., Shlens, J., & Sussillo, D. (2018). A dataset and architecture for visual reasoning with a working memory. In *European Conference on Computer Vision*, pp. 729–745. Springer.

Yang, S., Li, G., & Yu, Y. (2019). Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4145–4154.

Yang, Y., Teo, C. L., Daumé III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 444–454. Association for Computational Linguistics.

Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., & Salakhutdinov, R. R. (2016). Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pp. 2361–2369.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *arXiv preprint arXiv:1502.08029*.

Yao, T., Yingwei, P., Yehao, L., & Mei, T. (2017). Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*.

Yao, Y., Xu, J., Wang, F., & Xu, B. (2018). Cascaded mutual modulation for visual reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 975–980.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pp. 1039–1050.

Yin, X., & Ordonez, V. (2017). Obj2text: Generating visually descriptive language from object layouts. In *EMNLP*, pp. 177–187. Association for Computational Linguistics.

Yoshikawa, Y., Shigeto, Y., & Takeuchi, A. (2017). Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, *2*, 67–78.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4584–4593.

Yu, L., Bansal, M., & Berg, T. (2017). Hierarchically-attentive rnn for album summarization and storytelling. In *Empirical Methods in Natural Language Processing*.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., & Berg, T. L. (2018). Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315.

Yu, L., Park, E., Berg, A. C., & Berg, T. L. (2015). Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*.

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer.

Yu, L., Tan, H., Bansal, M., & Berg, T. L. (2017a). A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7282–7290.

Yu, Y., Ko, H., Choi, J., & Kim, G. (2017b). End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, pp. 3261–3269. IEEE Computer Society.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731.

Zeng, K.-H., Chen, T.-H., Chuang, C.-Y., Liao, Y.-H., Niebles, J. C., & Sun, M. (2017). Leveraging video descriptions to learn video question answering. In *AAAI*, pp. 4334–4340. AAAI Press.

Zeng, K.-H., Chen, T.-H., Niebles, J. C., & Sun, M. (2016). Generation for user generated videos. In *European conference on computer vision*, pp. 609–625. Springer.

Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S.-C. (2019). Raven: A dataset for relational and analogical visual reasoning. *arXiv preprint arXiv:1903.02741*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2018a). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H., Niu, Y., & Chang, S.-F. (2018b). Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4158–4166.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5014–5022.

Zhang, Y., Hare, J., & Prügel-Bennett, A. (2018a). Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*.

Zhang, Z., Xie, Y., & Yang, L. (2018b). Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6199–6208.

Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., & Qiao, Y. (2018). A multi-task learning approach for image captioning.. In *IJCAI*, pp. 1205–1211.

Zhao, Z., Yang, Q., Cai, D., He, X., & Zhuang, Y. (2017). Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, pp. 3518–3524. ijcai.org.

Zheng, Z., Wang, W., Qi, S., & Zhu, S.-C. (2019). Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6669–6678.

Zhou, L., Kalantidis, Y., Chen, X., Corso, J. J., & Rohrbach, M. (2018a). Grounded video description. *arXiv preprint arXiv:1812.06587*.

Zhou, L., Xu, C., & Corso, J. J. (2018b). Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, L., Zhou, Y., Corso, J. J., Socher, R., & Xiong, C. (2018c). End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8739–8748.

Zhou, M., Cheng, R., Lee, Y. J., & Yu, Z. (2018d). A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3643–3653.

Zhou, Y., Sun, Y., & Honavar, V. (2019). Improving image captioning by leveraging knowledge graphs. *arXiv preprint arXiv:1901.08942*.

Zhu, L., Xu, Z., Yang, Y., & Hauptmann, A. G. (2017). Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, *124*(3), 409–421.

Zhu, Y., Groth, O., Bernstein, M. S., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *CVPR*, pp. 4995–5004. IEEE Computer Society.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pp. 19–27. IEEE Computer Society.

Zhuang, B., Wu, Q., Shen, C., Reid, I., & van den Hengel, A. (2018). Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4252–4261.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pp. 391–405. Springer.