Multimodal Research in Vision and Language: A Review of Current and Emerging Trends

Shagun Uppal[a,1], Sarthak Bhagat[a,1], Devamanyu Hazarika[b,*], Navonil Majumder[a], Soujanya Poria[a], Roger Zimmermann[b] and Amir Zadeh[c]

[a] *Singapore University of Technology and Design, Singapore*
[b] *National University of Singapore, Singapore*
[c] *Carnegie Mellon University, USA*

ARTICLE INFO

Keywords:
Multimodal Learning,
Language, Vision,
Visual-Language Research,
Survey, Trends.

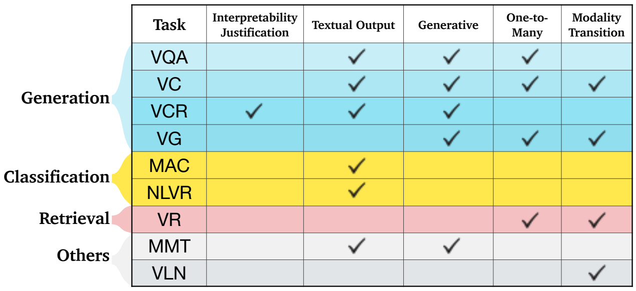
ABSTRACT

Deep Learning and its applications have cascaded impactful research and development with a diverse range of modalities present in the real-world data. More recently, this has enhanced research interests in the intersection of the Vision and Language arena with its numerous applications and fast-paced growth. In this paper, we present a detailed overview of the latest trends in research pertaining to visual and language modalities. We look at its applications in their task formulations and how to solve various problems related to semantic perception and content generation. We also address task-specific trends, along with their evaluation strategies and upcoming challenges. Moreover, we shed some light on multi-disciplinary patterns and insights that have emerged in the recent past, directing this field towards more modular and transparent intelligent systems. This survey identifies key trends gravitating recent literature in VisLang research and attempts to unearth directions that the field is heading towards.

1. Introduction

Computer Vision and Natural Language Processing have witnessed an impactful surge and development with the overall advancements in Artificial Intelligence. Independently, we have even surpassed human-level performance over tasks such as image classification, segmentation, object detection in vision and sentiment analysis, named-entity recognition in language research in supervised, unsupervised, and semi-supervised manners. With such powerful algorithms and comprehensive capabilities of autonomous systems comes the need to merge knowledge domains and achieve cross-modal compatibilities to innovate wholesome, intelligent systems. More often than not, we perceive real-world data and activities in multimodal forms involving multiple sources of information, especially at the intersection of vision and language. This has triggered Visual-Language (VisLang) research with more complex tasks and the need for interactive as well as interpretable systems. VisLang research has not only bridged the gap between discrete areas of interest, but also put forth the challenges and shortcomings of individual methods.

The integration of vision and language has been on various fronts through tasks such as *classification*, *generation*, *retrieval*, and *navigation*. This has surfaced various challenging tasks such as Vision-Language Navigation for the autonomous functioning of robots with a comprehensive understanding of its environment, Visual Captioning for generating rich and meaningful language descriptions from visual information, and many others. The field of Vision, Language, and VisLang research is undergoing rapid changes in trends and fast-paced progress. This makes it essential to bring together the recent trends and advances in VisLang research and take note of the current cutting-edge methodologies on multiple fronts. With this, we aim to identify and highlight the current challenges, critical gaps, and emerging directions of future scope that can help stimulate productive research.



	Task	Interpretability Justification	Textual Output	Generative	One-to-Many	Modality Transition
Generation	VQA		✓	✓	✓	
	VC		✓	✓	✓	✓
	VCR	✓	✓	✓		
	VG			✓	✓	✓
Classification	MAC		✓			
	NLVR		✓			
Retrieval	VR				✓	✓
Others	MMT		✓	✓		
	VLN					✓

Figure 1: A summary of VisLang tasks based on various underlying key characteristics.

Scope of the survey. This survey throws light on the fresh instigations in the sphere of VisLang research, enumerating the miscellaneous tasks that form the foundation of current multimodal research followed by the peculiar trends within each task. While prior studies have endeavored to perform similar analyses, our survey transcends them in task-specific and task-general inclinations in terms of architectures, learning procedures, and evaluation techniques. We also supplement our study with future challenges that lie in our path to developing self-sufficient VisLang systems that possess interpretable perception capabilities coupled with natural language apprehension, followed by future research direction in

[*] Corresponding author

shagun16088@iiitd.ac.in (S. Uppal); sarthak16189@iiitd.ac.in (S. Bhagat); hazarika@comp.nus.edu.sg (D. Hazarika); n.majumder.2009@gmail.com (N. Majumder); sporia@sutd.edu.sg (S. Poria); rogerz@comp.nus.edu.sg (R. Zimmermann); abagherz@cs.cmu.edu (A. Zadeh)
ORCID(s): 0000-0002-0241-7163 (D. Hazarika)

[1] Equal Contribution. Randomly Ordered.

this particular discipline.

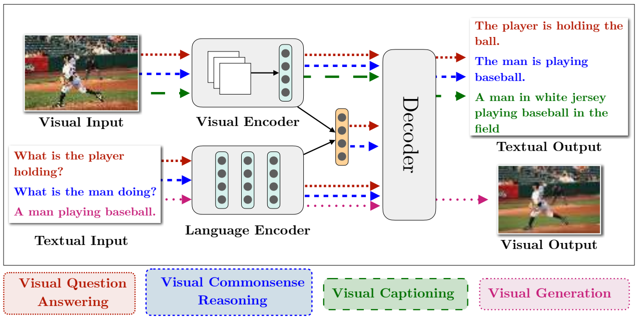
Related Surveys. Multiple recent works have delved into overviewing this field of research. These surveys provide an outline of numerous VisLang tasks and extensively detail the established datasets and metrics used for these tasks [81, 218]. Therefore, instead of focusing on similar attributes, we channel our attention to trends of specific tasks in terms of encoder-decoder architectures, attention mechanisms, learning techniques, amongst others. We also provide a brief overview of the foregoing VisLang metrics while describing the evolution of novel metrics and their significance in developing interpretable models with higher-order cognition capabilities, which prior surveys miss out.

Kafle et al. [145] points the readers towards the several challenges like dataset bias, robustness, and spurious correlations prevalent in VisLang research that could hinder their practical applications. While this work raises pertinent questions over current systems, it fails to invest in the contemporary evolution in these tasks that open new doors for eradication of these challenges. Additionally, Mei et al. [212] categorized VisLang tasks as either a transition from *vision to language* or *language to vision*. Under the bracket of *vision to language*, the authors provide detailed analysis of the task of visual captioning, providing insights into various encoder-decoder frameworks prevalent in such applications, while under *language to vision*, they recount the works that have focused on visual content creation. Although this work provides an intuitive element to VisLang tasks, it fails to consider other VisLang tasks that we illustrate and that together contribute to the development of extensive cognition and linguistic capabilities.

Baltrušaitis et al. [16] also provides a comprehensive overview of distillation and association of data from multiple modalities in terms of representation, translation, alignment, fusion, and co-learning trends. Our survey extends this abstraction by revealing the task-specific nature of these categorizations besides the ones overlooked by this work. Furthermore, while this survey covers a more general multimodal machine learning setting, we meticulously emphasize on VisLang tasks. Moreover, this work is a timely update of the latest trends in VisLang research, which have evolved more actively in the past two years.

Organization of the survey. In this survey, we begin by listing the diverse set of VisLang tasks alongside their mathematical problem formulations and categorization as per the fundamental problem at hand (Section 2). This is followed by a detailed overview of task-specific trends established in recent VisLang literature (Section 3). We also emphasize the trends regarding architectural formulation involving attention frameworks, transformer networks, muti-modal representation learning, and fusion techniques of the learned representations (Section 4). Furthermore, we demonstrate the evolving nature of imminent domains of interest in the VisLang community, including interpretability and explainability, multi-task learning, domain adaptation, and adversarial attacks. Lastly, we conclude with profuse challenges still prevalent in this active area of research, accompanied by the guidelines for future work gearing towards self-reliant VisLang systems (Section 6).

Figure 2: Overview of **Generation** Tasks.


Visual Input
Visual Encoder
Decoder
The player is holding the ball.
The man is playing baseball.
A man in white jersey playing baseball in the field
Textual Output
What is the player holding?
What is the man doing?
A man playing baseball.
Textual Input
Language Encoder
Visual Output
Visual Question Answering
Visual Commonsense Reasoning
Visual Captioning
Visual Generation

2. Tasks

A diverse range of tasks requires a coalesced and cooperative knowledge of both language and vision. Here, we discuss the fundamental details, goals, and trends of such tasks and how they have evolved in the recent past. Table 1 characterizes the various VisLang tasks on more fine-grained characteristics such as them being classification or generation problems, where there is a necessity for interpretable justifications, if the output is textual or not, and if a one-to-many mapping exists in an ideal and trivial sense. Modality transition distinctively refers to those tasks where the set of input and output modalities are disjoint, *i.e.* a given input in a particular modality needs to be represented in a completely different modality in the output space. Visual Question Answering, Visual Commonsense Reasoning, Visual Captioning, and Visual Generation correspond to generative models and methods, whereas the rest of them are majorly focused on perception tasks. Tasks like Visual Language Navigation improve upon the generalized as well as specific understanding of machines towards vision and language, without explicitly mapping none to the target space. We broadly categorize some of the tasks based on the underlying problem at hand, *i.e.*, *generation*, *classification*, *retrieval* or *others*.

2.1. Generation Tasks

We describe prominent generation tasks in VisLang, as illustrated in Figure 2.

Visual Question Answering (VQA) VQA represents the task of correctly providing an answer to a question given a visual input (image/video). For accurate performance, it is essential to infer logical entailments from the image (or video) based on the posed question.

For VQA, the dataset $\mathcal{D}$ generally consists of visual input-question-answer triplets wherein the i^{th} triplet is represented by $< \mathcal{I}_i, \mathcal{Q}_i, \mathcal{A}_i >$. We depict the set of all unique images by $\mathcal{V} = \{\mathcal{V}_j\}_{j=1}^{n_\mathcal{V}}$, set of all unique questions by $\mathcal{Q} = \{\mathcal{Q}_j\}_{j=1}^{n_\mathcal{Q}}$, and the set of all unique answers by $\mathcal{A} = \{\mathcal{A}_j\}_{j=1}^{n_A}$, where n_V, n_Q, and n_A represent the number of elements in these sets respectively. The core task involves learning a mapping function f that returns an answer for a given question with respect to the visual input, *i.e.*, $\hat{\mathcal{A}}_i = f(\mathcal{V}_i, \mathcal{Q}_i)$. The aim is to learn an optimal function f maximizing the likelihood between the original answers $\mathcal{A}$ and generated ones $\hat{\mathcal{A}}$. The output of the learnt mapping f could either belong to a set of possible answers in which case we refer this task format as *MCQ*, or could be arbitrary in nature depending on the question in which we can refer to as *free-form*. We regard the more generalized *free-form* VQA as a generation task, while *MCQ* VQA as a classification task where the model predicts the most suitable answer from a pool of choices.

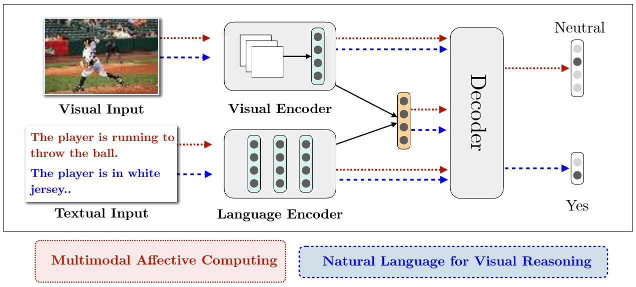
Visual Captioning (VC) Visual Captioning is the task of generating syntactically and semantically appropriate descriptions for a given visual (image or video) input in an automated fashion. Generating explanatory and relevant captions for a visual input requires not only a rich linguistic knowledge but also a coherent understanding of the entities, scenes, and their interactions present in the visual input.

Mathematically speaking, given a dataset $\mathcal{D} = \{< \mathcal{V}_1, \mathcal{C}_1 >, < \mathcal{V}_2, \mathcal{C}_2 >, ..., < \mathcal{V}_n, \mathcal{C}_n >\}$ with n data samples, the i^{th} datapoint $< \mathcal{V}_i, \mathcal{C}_i >$ represents the tuple of visual input $\mathcal{V}_i$ and its corresponding ground-truth caption $\mathcal{C}_i$. We learn a representation for the input to semantically encode the required information. The task is to use this information to generate a caption $\hat{\mathcal{C}}_i$ by maximising its likelihood with the ground-truth description. The generated description is a sequence of words (say k), is illustrated as $\mathcal{C}_i = \{c_i^1, c_i^2, c_i^3, ..., c_i^k\}$. Each token can be generated auto-regressively using sequential models, such as RNN or LSTM, based on the previous tokens.

Visual Commonsense Reasoning (VCR) Visual Commonsense Reasoning is the task of inferring cognitive understanding and commonsense information by a machine on seeing an image. It requires the machine to correctly answer questions posed about the image along with relevant justification.

Broadly, the task of VCR requires to learn a mapping from the input data distribution $\{< \mathcal{I}_1, \mathcal{Q}_1 >, < \mathcal{I}_2, \mathcal{Q}_2 >, ..., < \mathcal{I}_n, \mathcal{Q}_n >\}$, where $\mathcal{I}_i$ and $\mathcal{Q}_i$ depict the image and the corresponding query respectively, to the output comprising of answers and corresponding rationales namely, $\{< \mathcal{A}_i, \mathcal{R}_i >\}$. The rationales ensure that the right answers are obtained for the right reasons. The output distribution is commonly framed as answers to multiple-choice questions with explanations. Therefore, VCR can be broken down into a two-fold task that involves question answering (pick the best answer out of a pool of prospective answers to an MCQ question) and answer justification (provide a rationale behind the given correct answer).

Figure 3: Overview of **Classification** Tasks.


Visual Input
Visual Encoder
Decoder
Neutral
The player is running to throw the ball.
The player is in white jersey..
Textual Input
Language Encoder
Yes
Multimodal Affective Computing
Natural Language for Visual Reasoning

Natural Language for Visual Reasoning (NLVR). NLVR is a subtask of the broader category of VCR confining to the classification paradigm (as depicted in Figure 3). In a broader generalization, Natural Language for Visual Reasoning refers to the entailment problem, wherein the task is to determine whether a statement regarding the input image is *true* or *false*. The task formulation can be represented as learning a mapping from the input space comprising of images and queries, $\{< \mathcal{I}_1, \mathcal{Q}_1 >, < \mathcal{I}_2, \mathcal{Q}_2 >, ..., < \mathcal{I}_n, \mathcal{Q}_n >\}$ to the output space $< True, False >$ determining the truth value of an associated statement to each data point. It usually varies from VQA due to longer text sequences covering a diverse spectrum of language phenomenon.

Visual Generation (VG). Visual Generation is the task of generating visual output (image or video) from a given textual input. It often requires a sound understanding of the semantic information and accordingly generating relevant and context-rich coherent visual formations.

Given an input textual sequence of tokens, $\mathcal{T} = \{t_i^1, t_i^2, ..., t_i^k\}$, the aim is to output corresponding visual $\mathcal{V}$ capturing entities and scene illustrations as described in the text. It is a challenge to capture local and global context while synthesizing visualizations accurately. The output can be either an image or a video, based on various input forms being text descriptions, dialogues, or scene explanations.

2.2. Classification Tasks

We discuss specifics for classification tasks in VisLang, as illustrated in Figure 3.

Multimodal Affective Computing (MAC). Affective computing is the task of automated recognition of affective phenomenon causing or arising from emotions. Multimodal affective computing involves combining cues from multiple signals such as text, audio, video, and images depicting expressions, gestures, etc., in order to interpret the associated affective activity, similar to how humans explicate emotions.

Given a dataset $\mathcal{D} = \{< \mathcal{E}_1, \mathcal{T}_1 >, < \mathcal{E}_2, \mathcal{T}_2 >, ..., < \mathcal{E}_n, \mathcal{T}_n >\}$, where $\mathcal{E}_i$ and $\mathcal{T}_i$ denote a visual expression in the form of an image or video, and an associated text description respectively. Multimodal affective computing involves learning mappings from multimodal input signals to the decision space of different affective phenomena. Fusion of information from more than one signals to achieve consensus towards an emotion label provides human-level cognition and more reliable intelligent systems.

Figure 4: Overview of **Retrieval** Tasks.

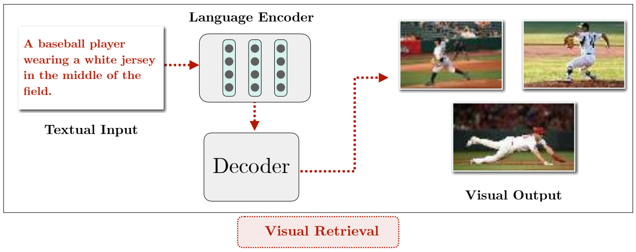
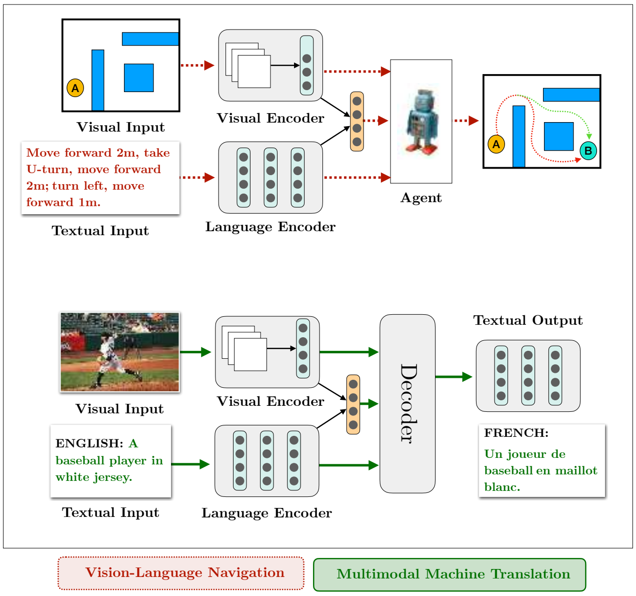

Language Encoder
A baseball player wearing a white jersey in the middle of the field.
Textual Input
Decoder
Visual Output
Visual Retrieval

Figure 5: Overview of **Other** Tasks.


A
Visual Input
Visual Encoder
Agent
A
B
Move forward 2m, take U-turn, move forward 2m; turn left, move forward 1m.
Textual Input
Language Encoder
Textual Output
Visual Input
Visual Encoder
Decoder
ENGLISH: A baseball player in white jersey.
Textual Input
Language Encoder
FRENCH: Un joueur de baseball en maillot blanc.
Vision-Language Navigation
Multimodal Machine Translation

2.3. Retrieval Tasks

We describe the task of Visual Retrieval, as illustrated in Figure 4.

***Visual Retrieval**.* The task of text-image retrieval is a cross-modal task involving the understanding of both language and vision domains with appropriate matching strategies. The aim is to fetch the top-most relevant visuals from a larger pool of visuals as per the text description.

Given a large database of n visual datapoints $\mathcal{D} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_n\}$ for any text description, say $\mathcal{T}$, we want to retrieve the top-most relevant images or videos from the database $\mathcal{D}$ as per $\mathcal{T}$. This is a cross-modal task due to text-based retrieval as opposed to other conventional approaches based on shape, texture, and color. This is popularly used in several search engines, domain-specific searches, and context-based image retrieval design systems.

2.4. Other Tasks

We describe the tasks of Vision-Language Navigation and Multimodal Machine Translation, as illustrated in Figure 5.

***Vision-Language Navigation (VLN)**.* Vision-Language Navigation is a grounding natural language task of an agent's locomotion as it sees and explores the real-world dynamics based on linguistic instructions. This is often viewed as a task of sequence-to-sequence transcoding, similar to VQA. However, there is a clear dichotomy between the two. VLN usually has much longer sequences, and the dynamics of the problem vary entirely because of it being a real-time evolving task.

Generally, for a given input sequence $\mathcal{L} = \{l_1, l_2, ..., l_n\}$, denoting an instruction with n tokens and o_1 representing the initial frame of reference, the agent aims to learn appropriate action sequences $\{a_1, a_2, ..., a_n\}$ following $\mathcal{L}$ to obtain the next frame of reference o_2 and continually so until the desired navigation task is complete. The key challenge lies in comprehending the environment and making confident decisions while exploring.

***Multimodal Machine Translation (MMT)**.* Multimodal Machine Translation is a two-fold task of *translation* and *description generation*. It involves translating a description from one language to another with additional information from other modalities, say video or audio.

Considering w.r.t to Visual Multimodal Machine Translation, we assume the additional modality as an image or a video. Given a dataset containing n data points, $\mathcal{D} = \{<\mathcal{V}_1, \mathcal{T}_1>, <\mathcal{V}_2, \mathcal{T}_2>, ..., <\mathcal{V}_n, \mathcal{T}_n>\}$, where $\mathcal{V}_i$ and $\mathcal{T}_i$ represent the visual input and the associated task description respectively, the aim is to learn a mapping to translated textual descriptions $\{\mathcal{T}'_1, \mathcal{T}'_2, ..., \mathcal{T}'_n\}$ in another language. The added input information is targeted to remove ambiguities that may arise in straightforward machine text translation and help retain the context of the text descriptions, considering the supplementary visual features. Multimodal representation spaces aid in robust latent representations complementing inherent semantic information held by visual and lingual embeddings individually.

3. Task-Specific Trends in VisLang Research

In this section, we look at latest papers published in concerned tasks and analyze emerging trends within the tasks. Figure 6 presents a rough estimate of the research trends across various VisLang tasks in the past two years. As seen in the figure, VC and VQA remain the most popular tasks. It is encouraging to see VCR emerge midway in the proportions suggesting an interest in the community towards reasoning-based tasks. We also provide a further breakdown of subtasks in VQA and VC in terms of subtasks that fall under them. While there have been a number of specific subtasks that have surfaced, typical tasks with images as their visual modality are most prevalent. The percentages depicted in figure are calculated based on the frequency of the papers published in these domains in recent literature.

3.1. Visual Captioning

***Image Captioning (IC)**.* Image Captioning [329, 48] comes under the multimodal visual captioning task wherein the in-

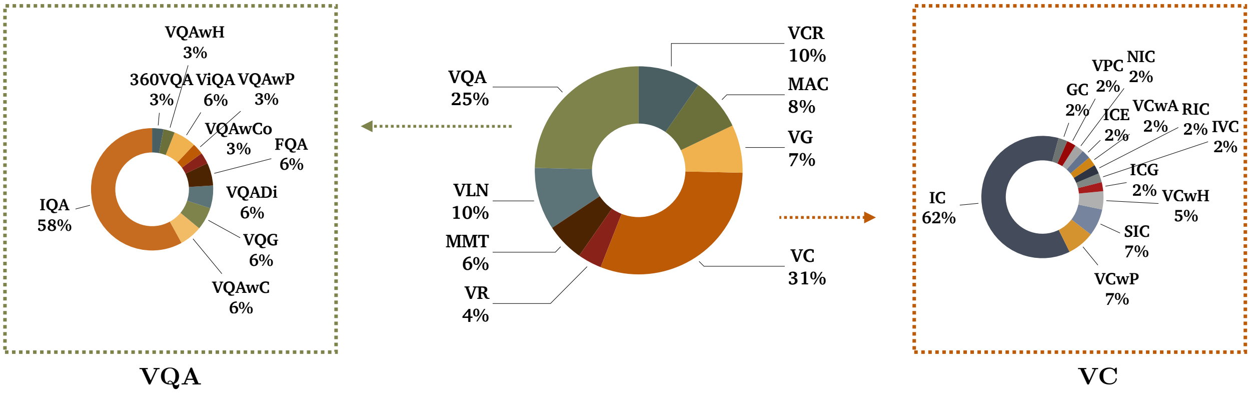


Figure 6: Paper trends of recent VisLang literature (previous 2 years). In this figure, we collate the task of NLVR with VCR due to its similar goals.

put to the model is an image. Recent advancements in the *Image Captioning* (IC) task have led to varied routes and applications for the same.

Images and captions can be correlated using relationship graphs for capturing underlying semantic information [281]. Such graphs can then be leveraged for generating novel captions in a weakly supervised setting. The identified relations are utilized to build coherence-aware models [6], capable of generating diverse captions based on different relation settings. Another IC task that has gained popularity is *Dense Image Captioning*. It involves generating multiple caption descriptions based on different potential regions in images [148]. Following up from previous paradigms is the task of *Relation-based Image Captioning* (RIC) wherein multiple caption descriptions are generated as per different relations identified amongst diverse regions in images [153].

Image Paragraph Captioning, as pursued by [161], generates detailed paragraphs describing the images at a finer level.

Video Captioning (VC). Another visual captioning application involves generating descriptions for video sequences [334, 38]. Analogous to *Dense Image Captioning* is the task of *Dense Video Captioning* [333, 410] wherein all the events in a video are described while generating captions.

Generating captions for videos in both sentence as well as paragraph formats has been pursued as a separate extended task, *Video Paragraph Captioning* (VPC) task [381].

Another approach referred to as *Audio-Visual Video Captioning* [311], combines multiple input modalities *i.e.* simultaneous audio and video signals to generate text-descriptions.

Others. The broad set of applications of IC has led to a diverse set of novel auxiliary tasks. One such task is *Visual Text Correction* which focuses on replacing the incorrect words in textual descriptions of videos or images with correct ones as per the visual content [211]. Similar tasks have been performed under *Image Caption Editing* (ICE) [269]. Other efforts have been made for tasks such as *Instructional Video Captioning* (IPC) [278]; captioning narrations of instructional videos, *Stylized Image Captioning* (SIC) [210]; for generating diverse image captions relating to specific semantic styles, *Image Captioning using Scene Graphs* (ICG) [45]; for captioning using scene graph representations of images, *Group Captioning* (GC) [186]; generating collaborative captions for a set of grouped images (albums) and *News Image Captioning* (NIC) [313]; generating captions for news article with associated images. Several other works highlight the visual captioning task with additional meta-data such as ratings and Part-of-Speech (POS), see Table 1.

Trends in VC. Visual captioning has been one of the most popular VisLang tasks that has gathered the attention of a wide-ranging community in developing models with strong cognition capabilities and intrinsic language understanding (see Table 1). Recently, an outburst of papers have focused on enhancing the perception capabilities of such models by bestowing supplemental sources of inference in the form of meta-data as illustrated in 3.1. *Stylized captioning* [210] is gaining momentum as an indispensable extension of VC that combines the idea of style-transfer (or feature swapping) from the vision domain with the setup of generating captions based on a visual input. Other vision-inspired ideas like representation learning and disentanglement have found immense range of applications in this field of research where developing robust representation can enhance the ability of a model to generate better captions. Despite the rapid developments in various attentional VC models, the bottom-up and top-down (up-down) attention [9] remains to be the most commonly applied framework in VC systems in present times, due to its ability to attend more naturally at the level of the objects and other principal regions.

A large portion of visual captioning approaches that tend to form individual encodings for visual and language input deploy an object-detection network. This network identifies the set of all possible entities present in the visual input and draws correlation between the ones identified by the language models in the textual input. In recent works, the most predominantly utilized object detection model was Faster RCNN [93] followed by YOLO [261] and RCNN [93] owing to its highly accurate predictions coupled with fast computations. Alternative approaches that do not employ object detection models

Table 1
Latest research in IC

Ref.	Task (Dataset)	Visual Encoder	Language Model	Attention
[281]	IC (MSCOCO)	Faster RCNN	Graph Parser/LSTM	Soft, Adaptive
[6]	SIC (Clue)	ResNet-50	GloVe	Multi-head
[171]	VPC (ActivityNet Captions, YouCookII)	CNN	Transformer	Multi-head
[103]	IC (MSCOCO)	Faster RCNN	GloVe, LSTM	Self
[400]	VC (MSVD, MSR-VTT, VATEX)	2DCNN, 3DCNN, GCN	LSTM	Temporal, Spatial
[45]	ICG (VisG, MSCOCO)	MR-GCN	LSTM	Graph-based
[405]	SIC (SentiCap and FlickrStyle10K)	Scene Graph	LSTM	Top-down, Visual
[186]	GC (Conceptual Captions, Stock Captions)	ResNet50	LSTM	Self
[412]	IC (Flickr30k, MSCOCO)	Faster RCNN	LSTM, GRU	Up-Down
[269]	ICE (MSCOCO)	RCNN	LSTM	SCMA
[60]	IC (MSCOCO)	Faster RCNN	GloVe	Cross, Self
[313]	NIC (NYTimes800k, GoodNews)	ResNet-152, MTCNN, YOLOv3	RoBERTa	Multi-head
[229]	IC (MSR-VTT, MSVD)	ResNet-101, I3D, Faster RCNN	Transformer, GCN	Temporal, Spatial
[335]	IC (MSCOCO)	RCNN	LSTM	Recalled-words
[273]	ICwH (Conceptual Captions, Caption-Quality, Conceptual Captions Challenge T2)	Faster RCNN, Google Cloud Vision API	BERT	-
[403]	IC (Conceptual Captions)	CNN, Transformer	Transformer	Self
[278]	IVC (YouCookII)	ResNet-34, Transformer	LSTM, BERT	Self
[77]	IC (MSCOCO, Flickr30k)	ResNet-152	RNN	-
[130]	IC (MSCOCO)	Faster RCNN	LSTM	Base, Recurrent, Adaptive
[114]	IC (MSCOCO)	Faster RCNN	Transformer	Self
[37]	IC (MSCOCO)	CNN	LSTM	-
[331]	VCwP (MSR-VTT, MSVD)	CNN	LSTM	Soft
[123]	VCwP (MSR-VTT, MSVD, ActivityNet)	CNN	ConvCap	Soft
[91]	IC (MSCOCO)	CNN	MaBi-LSTM	Cross-modal
[375]	IC (MSCOCO)	Faster RCNN, Mask RCNN, Tree-LSTM, GCN-LSTM	LSTM	Up-down
[193]	ICwP (MSCOCO)	Faster RCNN	LSTM	Textual, Visual
[367]	IC (MSCOCO, VisG)	CNN, Faster RCNN	RNN	Object, Attribute, Relation, Self
[12]	IC (MSCOCO)	Sequencial VAE	LSTM	-
[166]	IC (MSCOCO, Flickr30k, Conceptual Captions)	CNN	GRU, GloVe	-
[150]	IC (MSCOCO)	Faster RCNN	LSTM	Reflective
[326]	IC (MSCOCO)	ResNet, Faster RCNN	SCG	-
[257]	ICwA (ActivityNet)	3DCNN, C3D	GRU RNN	Crossing
[99]	IC (VisG, MSCOCO)	Faster RCNN	LSTM	Graph
[277]	ICwH** (MSCOCO)	CNN	RNN, CNN	Up-down
[366]	IC (MSCOCO)	CNN	RNN	Up-down
[67]	ICwP (MSCOCO)	VGG-16, Faster RCNN	LSTM, CNN	-
[80]	IC (MSCOCO)	CNN	LSTM	-
[377]	IC (VisG)	Faster RCNN	LSTM	-
[153]	RIC (VisG Relationship v1.2)	VGG16	LSTM	-
[59]	IC (Flickr30k, MSCOCO)	Faster RCNN	LSTM	Adaptive
[88]	IC (MSCOCO)	CNN	RNN, LSTM	Up-down
[256]	IC (VisG, MSCOCO)	Faster RCNN	LSTM	Look Back
[408]	IC (MSCOCO, ImageNet)	CNN	LSTM	Up-down
[72]	IC (OOC, MSCOCO)	ResNet-101	LSTM	Co, Visual, Context-aware
[329]	IC (Pascal, MSCOCO, Flickr30k)	CNN	RNN	-
[359]	IC (Flickr9k, Flickr30k, MSCOCO)	CNN	RNN	Soft, Hard
[374]	IC (MSCOCO, ImageNet)	FCN, VGG-16	LSTM	-
[210]	SIC (MSCOCO, Styled Text)	CNN	GRU	Up-down
[46]	SIC (MSCOCO, FlickrStyle10k)	CNN	LSTM	Style, Self
[209]	SIC (MSCOCO)	VGG CNN, RNN	RNN, LSTM	-
[282]	SIC (Flickr30k, MSCOCO)	ResNet-152, FC	Transformer, FC	Up-down
[265]	IC (MSCOCO)	ResNet-101, FC	LSTM	Hard
[373]	IC (MSCOCO)	CNN, RNN	RNN	Hard, Soft

for encapsulating visual object entities, often utilize a deep convolution network like various variants of ResNet [112] (16, 50, 101 or 152) or VGGNet [284] (16) in order to encode the image onto a lower-dimensional manifold for extraction of relevant visual features.

Visual captioning, being one of the most primitive fields in the VisLang domain, has a wide variety of benchmark datasets. The most commonly utilized dataset for this particular task remains MSCOCO [192] that consists of images of complex scenes with common everyday objects in their natural context. However, due to the numerous concerns raised over the interpretability of popular VC models, novel datasets that require complex reasoning and cognition capabilities have arisen in recent literature that have contributed significantly to the transparency, fairness, and explainability of VC systems (refer Section 5.2).

3.2. Visual Question Answering. (VQA)

The VQA task has had a history of diverse applications and proposed architectures to achieve them.

Image Question Answering (IQA). The IQA task requires inferring semantic and abstract concepts in images to perform question-answering with the acquired knowledge with high fidelity. Several techniques have been explored to achieve benchmark results for IQA with attention-based methods emphasizing focus on the vital features. This is achieved either using a co-attention model for VQA with attention over both image and question inputs to answer *where to look?* and *what to look for?* simultaneously [198], or using stacked attention networks in order to infer correct answers using multiple queries progressively [373]. *Knowledge-based* methods [352] have also been explored where external knowledge apart from the content available in the image is utilized so as to understand the scene representations deeply and to be able to answer a wide and deep set of related real-world questions. *Memory networks*, which leverages attention mechanisms along with memory modules to achieve benchmark performances, and can be generalized to other modalities apart from images, such as text, have also been used for the task of VQA [356]. Another dimension to IQA involves interpreting figures, plots, and visualizations and answering relevant questions based on the data visualization [36, 144], referred to as *Figure Question Answering*. Attention techniques and bimodal embeddings have commonly been used to infer plots and charts as inputs. Chou et al. [58] introduced a novel task of 360° *Visual Question Answering (*360*VQA)* wherein the inputs are images with a 360° field of view. Such an input representation provides a complete scenic understanding of the entities in the image but also demands models with spatially-sound reasoning abilities to leverage the extra information available.

Video Question Answering (ViQA). ViQA involves answering questions based on temporal data in the form of video sequences. With the limited availability of annotated resources, this has been pursued using online videos along with available descriptions to obtain question-answer (q&a) pairs in an automated fashion instead of manual annotations [390]. These q&a pairs are used in training the ViQA model. Tapaswi et al. [307] introduced a novel MovieQA dataset on similar lines, to perform question-answering on the events occurring in movie videos. It aims to comprehend a multimodal video-text input signal for visual question-answering tasks. Apart from free-form answers, several efforts have been made to attain *fill-in-the-blank* type of inference over the knowledge of events described in video inputs [414]. It used recurrent neural networks in an encoder-decoder setting.

Visual Question Generation (VQG). The VQG task requires generating natural questions given the images. It demands a more intensive subject-capturing of the context to generate a relevant and diverse set of questions such as ones with answer categories belonging to a specific bracket like spatial, count, object, color, attribute etc. Whereas other tasks like IQA, tend output answers on a broader level like binary answers to questions. Informative generations have been synthesized using structured constraints such as triplet loss with multimodal features [237]. Other architectural variants like VAE-LSTM hybrids help synthesize largeset of questions, for a given image [138], or in a dual task setting of question-answering [368, 182]. A paradigm translation has been posed from neural network-based approaches to reinforcement learning settings [397] for VQG task, with optimal rewards based on informativeness of generated questions.

Visual Dialog Visual dialogue comprises of *VQA in dialogue (VQADi).* and *VQG in dialogue (VQGDi)* wherein the main goal is to automate machine conversations about images with humans. Probabilistic approaches [236] have been undertaken to ensure minimum uncertainty of generated dialogues given the history of the conversation. Similar to other vision-language tasks, attention-based methods [226, 147, 272] have helped capture multimodal references in the dialogue stream. Dialog systems have leveraged generative modeling [137, 208] using both question and answering tasks coherently. Moreover, Das et al. [65] and Zhang et al. [398] proposed the Visual Dialog task in a deep reinforcement learning setup using co-operative agents.

Others. A variety of other task extensions with specific areas of interest have emerged from VQA. As a hybrid of VQA and dialogue systems, *VQA in Dialogue* (VQADi) and *VQG in Dialogue* (VQGDi) have surfaced interest [170, 101]. As a broader scale application, VQAwCo poses the problem of visual question answering based on a collection of videos or photos [187]. Other variants involve leveraging subsidiary information in the form of metadata such as VQA with paragraph descriptions (VQAwP) [154], image captions (VQAwC) [409] and human visual or textual inputs (VQAwH) [350].

Trends in VQA. The task of VQA has grown by several folds in the past decade wherein the development of attention frameworks promoting enhanced comprehension proficiencies have been a fundamental component of modern VQA systems. As a result, explainability-motivated frameworks like hierarchical and graph-based attentions have found great applications in VisLang research. Co-attention has recently evolved as the most commonly deployed attention framework due to its potential of associating key objects in the visual inputs (identified via image encoders described in Section 4.1) to textual entities in the question. Owing to the recent developments in various language generation tasks, the VQA community is drifting toward a generalized form of free-form question-answering, while diverging from the simpler form of MCQ answers. While VQA v2 [2] remains to be the repeatedly operated benchmark dataset, an abundant set of datasets that focus on very specific sub-tasks of VQA like 360° images VQA [58], visual dialogue question answering [64], VQA by reading text in images *i.e.* by optical character recognition (OCR) [215], *etc.* have come to light.

In terms of the visual encoder of VQA systems, similar to the task of VC, object detection models are common in identifying the essential entities in the input and correlating them with the ones obtained in the language input. A number of approaches have also tried to maintain the simplicity in

Table 2
Latest research in VQA

Ref.	Task (Dataset)	Visual Encoder	Language Model	Task Format
[369]	ViQA (TVQA, Pororo)	Faster RCNN, BERT	BERT	MCQ
[57]	360VQA (360° VQA Dataset)	CNN	GRU	MCQ
[235]	IQA (VQA-X Dataset)	CNN	LSTM	MCQ
[36]	FQA (LEAF-QA, FigureQA, DVQA)	MaskRCNN, Oracle/OCR	LSTM	MCQ
[144]	FQA (FigureQA, DVQA)	CNN, FC	LSTM	MCQ
[409]	VQAwC (MSCOCO, Flickr30k, VQA 2.0)	Faster RCNN	Transformer, BERT	MCQ
[90]	ViQA (KnowIT VQA)	ResNet-50, OD, FRN	BERT	MCQ
[348]	VQAwC (VQA 2.0)	CNN	GRU	MCQ
[131]	IQA (VQA 2.0)	Faster RCNN	GloVe, ELMo	MCQ
[154]	VQAwP (VisG Dataset)	Faster RCNN	LSTM	MCQ
[71]	IQA (TDIUC, VQA2.0, Visual7W)	FPN Detector	GRU	MCQ
[89]	IQA (VQA 2.0, TDIUC)	RCNN	Transformer	MCQ
[174]	IQA (VQA 2.0, VQA-CP v2)	Faster RCNN	Bi-RNN, GRU	MCQ
[21]	IQA (VQA 2.0, VizWiz)	CNN	GRU	MCQ
[349]	VQAwH (VQA-CP v2)	Faster RCNN	GRU	MCQ
[384]	IQA (VQA 2.0)	Faster RCNN	LSTM, GloVe+	MCQ
[162]	VQG (VQA 2.0)	CNN	LSTM	Free Form
[320]	VQG (VQA 2.0)	CNN	LSTM	Free Form
[305]	IQA (VisG, VQA 2.0)	Visual Attention Module	Bi-TreeLSTM	MCQ
[24]	IQA (VQA 2.0, VQA-CP v2, TDIUC)	Faster RCNN	GRU	MCQ
[101]	VQADi (Visual Dialogue v1)	CNN	LSTM	Free Form
[40]	IQA (VQA-CP v2)	Faster RCNN	LSTM	MCQ
[170]	VQGDi (GuessWhich Task)		RNN	Free Form
[9]	IQA (VQA)	Faster RCNN	LSTM	MCQ
[84]	IQA (VQA, Visual7W)	CNN	WE, LSTM	MCQ
[198]	IQA (VQA, COCO-QA)	VGGNet/ResNet	LSTM	MCQ
[151]	IQA (MM IMDB, FOOD101, V-SLNI)	ResNet, CNN	Bi-Transformer	MCQ
[358]	IQA (DAQUAR, VQA)	GoogleLeNet	Word Embeddings	MCQ
[371]	IQA (DAQUAR-ALL, DAQUAR-REDUCED, COCO-QA, VQA)	CNN	CNN/LSTM	MCQ
[187]	VQAwCo (MemexQA, MovieQA)	CNN	LSTM	MCQ
[312]	IQA (Visual Madlibs)	VGG-16, Faster RCNN, SSD	word2vec	MCQ

their architecture by using fundamental convolutional layers with or without fully connected layers at the end of it for encoding the visual inputs [21, 162, 320, 144]. Pre-trained transformer-based embeddings have certainly boosted the performance of models that aim at generating individual embeddings for each modality that are later fused to obtain a hybrid latent code.

Table 2 depicts the relevant recent literature that have focused on the application of VQA using diverse visual encoders and language models.

3.3. Visual Commonsense Reasoning (VCR)

The task of VCR [389] was introduced to develop higher-order cognition in vision systems and commonsense reasoning of the world so that they can provide justifications to their answers. Encodings produced by BERT-inspired transformers models proved to implicitly establish relationships between entities present in the multimodal sources of data aiding the process of reasoning. BERT embeddings were directly used [158] guided by attention for the task of VCR. Multimodal extensions of BERT like ViLBERT [196], VisualBERT [176] and VL-BERT [291] also justified the efficacy of such embeddings in developing sophisticated understanding required for reasoning over a wide domain of questions. On similar lines, the joint vision and text embeddings space was learnt using large-scale pre-training [49] for a variety of multiplex multimodal tasks, including VCR. For constructing such joint representations that capture complex relationships between objects present in the visual scene, knowledge from scene graphs was incorporated while pre-training their model [380]. Adversarial training on multiple modalities for developing these joint embeddings spaces has also been studied in [85]. Their model was trained via a two-step framework that includes task-agnostic pre-training, which is followed by task-specific fine-tuning.

Trends in VCR. The growing focus on reasoning based systems have not only led to instigation of various relevant datasets but also popularized the idea of visual reasoning (see Table 3). VCR [389], often viewed through the lens of an extension of the VQA task, is slowly and steadily gaining extensive popularity in recent literature owing to its contribution in designing socially relevant interpretable models. Most VCR systems have focused on utilizing embeddings generated from pre-trained transformer networks (detailed in Section 4.1) to obtain high-quality latent representations for visual and language modalities. This evolution has also prompted casuality, counterfactual inference and contrastive learning to flourish in these domains of interest.

The subtask of VCR, NLVR usually varies from VQA due to longer text sequences covering a diverse spectrum of language phenomenon. Most common visual reasoning approaches target at mimicking the human brain cognition of identifying the broader concepts in the visual input, making it easier to apprehend implicit relationships between different entities with respect to these concepts. The recently intro-

Table 3
Latest research in Visual Commonsense Reasoning. CTM: Contextual Voting Module, CC: Conceptual Captions, KV: Key Value

Ref.	Dataset	Visual Encoder	Language Model
[291]	CC, VCR, VQA 2.0, RefCOCO+	Faster RCNN, ResNet-101	Transformer
[407]	CLEVR, NLVR	CNN, GRU	KV Mem. Net
[406]	VQAv2, NLVR	Faster RCNN	BERT
[172]	CC, VCR, MSCOCO, Flickr30k	Faster RCNN	Transformer
[110]	COG	-	1DCNN/LSTM
[24]	VQA 2.0, VQA-CP v2, TDIUC	Faster RCNN	GRU
[279]	CLEVR	TbD-net	LSTM
[382]	VCR	CTM, ResNet-50	BERT
[191]	VCR	CNN	BERT
[347]	VCR	CNN, GCN	BERT, LSTM
[196]	VQA 2.0, RefCOCO+, VCR, Flickr30k	Faster RCNN	Transformer, Co-TRM
[49]	MSCOCO, CC, VisG, SBU Captions, Flickr-30k, VQA 2.0, NLVR, RefCOCO+	RCNN, FC	Transformer
[389]	VCR	CNN	BERT

duced NLVR task [293] focuses on how visual theoretic reasoning can be translated to answering multiple linguistic phenomenon. This has been further explored with bidirectional matching [302] benefitting in end-to-end frameworks with attention based settings. Such approaches have been extended to more generalized settings to recognize unseen object images as well [106]. This generalization is achieved via learning surplus meta-concept learners capturing two-way relationships between visual concepts and meta-concepts (for *e.g.*, properties of objects like color and shape). Other supplementary NLVR settings involve using descriptions for pair-level images as input [294].

3.4. Multimodal Machine Translation (MMT)

MMT is a task wherein visual data acts as a supplement for fostering the primary task of translating descriptions from one language to another. Related works [26] emphasize the fact that multimodal sources of data tend to enhance the performance of a model when performing machine translation. Using this principle, attention-based approaches [133] have been proposed for generating informative multimodal embeddings that could be translated by the decoder. Inspired by the success of multimodal attention for IC, other methods [27] utilize it for MMT in order to simultaneously focus on the image and text description. This followed a diverse pool of Seq2Seq models with attention mechanisms to be introduced for this task [189, 411, 29].

Trends in MMT. MMT is one of the most primitive VisLang tasks that has been a source of interest in the community for a long time, but it has gained escalation in several recent works focussing on this task (see Table 4). The emergence of generalizable BERT-based pre-trained latent spaces has led to a boost in the performance of MMT models over a wide range of global and regional languages. Similar to VC, former datasets like Multi30k [75] are still the most popular benchmarks for this particular task. However, the growth of reasoning-oriented models has led to some newer datasets (such as VATEX [341] and Flickr30-Entities [244]) that demand higher-level capabilities.

Table 4
Latest research in Multimodal Machine Translation.

Ref.	Dataset	Visual Encoder	Language Model
[118]	Multi30K	ResNet-50	Bi-GRU, GloVe
[365]	Multi30K	Faster RCNN	RNN
[132]	Multi30K	MLP	Transformer
[30]	Multi30K, Comp. Multi30K	FC	Bi-LSTM
[136]	Multi30K	ResNet-50	GloVe
[44]	IAPR-TC12, Multi30K	ResNet-152	BiLSTM
[292]	Multi30K	ResNet-152	Transformer
[119]	Multi30K	ResNet-50	Bi-GRU

3.5. Multimodal Affective Computing

Multimodal affective computing comes in intuitive to attain human-level accuracy since humans comprehend varied emotions with an integrated knowledge of sound, visual expressions and the semantic context of lingual information. More recently, text and images have been combined to infer the associated sentiments more adequately [139], similar to how most social media platforms allow and access information. A lot of approaches [52, 249] leverage facial expressions combined with audio or text to learn correlations between different information types for a fine-grained emotion classification. A great deal of attention has also been given to the different possibilities of fusing multimodal information into purposeful representations. Fusion of multiple sensory data into a single information channel at the feature-level [219, 270, 248] has proven to provide high-fidelity classification for a diverse set of underlying tasks. Several methods explore the Hidden Markov Models (HMMs) to model varying levels of correlation in different signal input for fusion [391, 289].

Diverging from trivial sentiment analysis based on textual sequences, multimodal affective computing targets on using manifold cues from differing input signals like visual, audio or text [251], be it using facial expressions or gait and gestures or speech (vocal) features. Considering this, several works have touched upon different sub-aspects of MAC:

Multimodal Sentiment Analysis. Multimodal Sentiment Analysis majorly focuses on broadly measuring the emotion on the extremity scales such as *positive*, *negative*, *neutral* instead of more fine-grained classification based on precise emotions and opinions. Taking into consideration the interdependencies in multiple utterances [247] of a video along with multi-kernel learning [246] helps improving performance on sentiment classification objectives. Also, a varied level of strategies are adopted for heterogenous modality fusion mechanisms. Some of the non-conventional fusion methods which obtain a boost in the performance, use either hierarchical fusion, combining only two modalities at any individual level [205], or using a tensor-fusion network by blending different modality representations at a deeper layer in contrast to early fusion [386] for effectively encapsulating inter-modal as well as intra-model correlations. Zadeh et al. [388] introduced a large multimodal emotion recognition and sentiment analysis dataset along with another hierarchical fusion technique involving dynamic fusion graphs operating on different degrees of freedom at each level. For multi-view sequence learning, with differing modalities as varying views, Zadeh et al. [387] used gated network along with attention model to apprehend the heterogeneity.

Affective Computing with Vision-Text. A combination of textual and visual information is often popularly observed on social media platforms which usually sees a large incoming flow of data. To assess various such data in the form of tweets or other social media data, Cai and Xia [28] proposed convolution based networks for separately encoding unimodal informations and then amalgamating them through another convolution model. To effectively capture the interdependencies in heterogenous modality spaces, probabilistic graphical models were employed, along with hyper-graph models for analysing independent features [140].

Affective Computing with Vision-Text-Audio. Trimodal features involving visual, textual and acoustic features obtain optimal performance as compared to individual information signals. Achieving modality-invariant features with appropriate fusion techniques plays a significant role in using available information aptly. Fusion could be achieved with either single kernel [241] or multiple kernel learning [250]. Using video data allows to extract three modalities from a single source of data [248, 283], but at the same time poses an additional challenge of coherently extracting and segregating different modalities before processing.

Trends in MAC. The latest trends in multimodal affective computing span several promising directions (see Table 5). One class of models improve upon tensor-based fusion methods and attempt to find efficient solutions to otherwise inefficient process [195, 18]. Recent works, such as Liang et al. [188] also address parallel, yet significant issues that include accounting for temporal imperfections in multimodal time-series data.

Another popular research line is in multimodal representation learning, which is often either a replacement or a precursor to multimodal fusion. The former is predominantly observed in modality-translation based methods that have the attractive property of robustness against missing modalities [243]. Pushing the goal towards effective multimodal representations are works like [316, 111] that attempt to factorize or disentangle modality features in joint spaces. Another exciting direction is in learning alignment-independent methods for multimodal fusion, which alleviates the labor-intensive process of cross-modal alignment in the ground truth annotations [317].

Table 5
Latest research in Multimodal Affective Computing (MAC). Note: FACET is available at `https://imotions.com/platform/`.

Ref.	Dataset	Visual Encoder	Language Model
[111]	MOSI, MOSEI, UR_FUNNY	Facet, LSTM	GloVe, BERT
[298]	MOSI, MOSEI, IEMOCAP	Facet	BERT
[204]	MOSI, MOSEI, IEMOCAP	Facet, 3d-CNN	GLoVe, CNN
[18]	MOSI, POM, IEMOCAP	Facet, LSTM	GLoVe, LSTM
[317]	MOSI, MOSEI, IEMOCAP	Facet, Transformer	GLoVe, Transformer
[344]	MOSI, IEMOCAP	Facet, LSTM	GloVe
[316]	POM, MOSI, ICT-MMO MOUD, Youtube, IEMOCAP, SVHN, MNIST	Facet, CNN, LSTM	GloVe
[202]	MOSI, MOSEI, IEMOCAP	Facet, 3d-CNN	GLoVe, CNN
[243]	MOSI, ICT-MMO, Youtube	Facet	GLoVe
[203]	MOSI, MOSEI, IEMOCAP	Facet, 3d-CNN	GLoVe, CNN
[109]	UR_FUNNY	OpenFace, LSTM	GLoVe, LSTM

Table 6
Latest research in Vision Language Navigation (VLN)

Ref.	Dataset	Vision Model	Language Model
[122]	FGR2R	ResNet-152	LSTM
[379]	R2R	Seq2Seq	-
[418]	CVDN	CNN	LSTM
[354]	R2R	CNN	LSTM
[339]	R2R	CNN, LSTM	LSTM
[108]	R2R, CVDN, HANNA	ResNet	Transformer
[53]	R2R	ResNet	BiLSTM
[167]	R2R	ResNet-152	Pos. Encoding
[413]	R2R	LSTM	BiLSTM
[83]	R2R	LSTM	LSTM
[179]	R2R	LSTM	LSTM, GPT
[128]	R2R	CNN	BiLSTM
[168]	R2R	ResNet-152	LSTM
[303]	Matterport3D	LSTM	BiLSTM
[149]	R2R	Seq2Seq	LSTM
[200]	R2R	ResNet-152	LSTM

3.6. Vision-Language Navigation (VLN)

VLN [10] recently emerged from combining separate visual-based [214] and language-based [11] navigation tasks. Several methods have either used self-supervised [338] or self-correcting [149, 201] strategies to improve path planning for navigation. Most of the successful attempts at solving the task of VLN have been inspired by reinforcement learning [342] and imitation learning [338] in contrast to the earliest Seq2Seq models [10]. Contrary to discrete settings, novel approaches focus on improving real-time navigation in continuous domain 3D environments [160] by identifying novel objects which were unseen before [254].

Trends in Vision-Language Navigation. With the dynamic nature of the task of navigation using language instruction, a wide variety of works [342, 173] have resorted to utilizing (deep) reinforcement learning techniques for fabricating adaptive generalizable models for a range of environmental settings. In such approaches, the agent learns to generate a map of the environment alongside following the instructions to advance towards the goal by receiving rewards from the environment. Even contemporary learning-based techniques like imitation learning have been popularized to learn to navigate using instructions by mimicking an instructor's actions. Despite the rapid growth in VLN, the conventional Room-to-Room (R2R) dataset [10] remains to be the most sought after benchmark in this task. Table 6 shows the recent trends in the VLN research.

3.7. Visual Generation

Visual generation has been carried under different task variations as highlighted under.

Text-to-Image Generation (T2I). The task of generating high fidelity images from textual descriptions [263] has been accomplished by utilizing GANs to capture visual concepts from textual descriptions and then translating them onto images. StackGANs [395] disintegrated this task of synthesizing images from descriptions into disjoint steps that first apprehended the rudimentary concepts of the image like shape and color to form low-resolution images, which were later used to generate high-resolution images with finer details. To its further extension, tree-structures are introduced with multiple generators and discriminators Zhang et al. [396] arranged together. Utilizing this tree of networks, images with varying scales are generated from different tree branches in an unconditional or conditional setting. A fine-grained image-text matching loss combined with a multimodal attentional GAN architecture, conditions on given text at word level to generate high-quality images [360]. Additionally, hierarchical networks with hierarchical-nested adversarial objective were proven to aid generator training, forming high-resolution photographic images [401].

Dialogue-to-Image Generation (D2I). Utilizing dialogues as a supplemental source of contextual information for the generation of images can lead to the fabrication of meaningful real-looking images. This leads to some novelty in the task of text-to-image generation by additionally utilizing dialogues for encapsulating finer details to improve image-generation [275].

Another active area of research *dialogue-to-series-of-images generation* (D2SI) seeks to generate a sequence of images rather than a single one iteratively, making use of sequentially appearing texts or feedbacks. This task requires a deeper understanding of the context and entities detailed in the text acquired from the previous image output and all preceding feedbacks [74].

Agent-guided dialogue-to-scene generation (AG-D2S), another subcategory of D2I, is the task of designing an entire scene using multi-agent collaboration by interacting with other entities in the input. Generally, the task is solved by a collaborative scene construction between two agents wherein one of them instructs the other by capturing key semantic and contextual ideas, while the other draws the scene on the empty canvas [155, 378].

Dialogue-based image editing (DIE) [51] is another task that aims at sequentially editing images based on the textual instructions provided by the user, improving the quality of the image at each step in the process. The model is required to maintain consistency between the user descriptions provided and the generated image besides simultaneously modifying it region-wise in an iterative manner.

Scene Graph/Layout-to-Image Generation. The task of generating real-world images based on textual instructions about individual objects and their locations in the image has been performed by training a GAN conditioned over both the descriptions as well as the object locations [262]. Some approaches have also tried to breakdown this process into a sequence of similar steps. This includes first generating the overall layout using textual descriptions, and then generating the images using a separate generator. This generator that synthesizes the images in a coarse-to-fine manner by generating bounding boxes for all objects and then refining each of the objects in them sequentially [121]. Further extensions to natural language descriptions involve rendering scene graphs for video synthesis to efficiently capture entity-object relationships on more complex domains [142] or object-box layouts [402]. Devoid of GANs, certain approaches generate scene objects sequentially by attending on previous state of the generated scene dynamics [301].

Text-to-Video Generation. Generating videos from textual descriptions enhances the challenge a level up than image generation for deep generative models due to temporal nature of output and more variable dynamics. To effectively capture and synthesize features with differing frequencies in a video, conditional generation helps to segregate static and dynamic features from text [183]. Another extension to standard GANs modifies discriminator networks to verify generated video sequences against correct captions instead of real/fake, with spatio-temporal convolutions for synthesizing frames [231]. Hybrid models with variational-recurrent attention mechanisms also demonstrate high-fidelity generations [216] with individual frames attended, using LSTMs for video frame predictions [43].

Trends in VG. With the rapid advancements in GAN-based architectures eliciting the high fidelity visual generation, the task of VG has expanded by many folds in the recent literature (see Table 7). Under the envelop of VG, several particular subtasks like generation of scenes guided by dialogue or human feedback have surfaced. Whilst principally most approaches utilize GANs for the generation task, recent furtherance in VAE-based [321] or flow-based [156] probabilistic generative models have provided extraordinarily fine details in visual outputs. These trends open new doors for VG research to generate high-quality images or videos avoiding the pitfalls of GAN's training stability issues.

Table 7
Latest research in Visual Generation. ConvRM: Convolutional Recurrent Module

Ref.	Dataset	Visual Generator	Language Model
[376]	CUB, MSCOCO	GAN	Bi-LSTM
[178]	MSCOCO	GAN	Bi-LSTM, LSTM
[255]	CUB, MSCOCO	GAN	RNN
[207]	CIFAR-10, CUB, facades, maps, Yosemite, cat–dog	GAN	LSTM
[416]	Caltech-UCSD Birds200, MSCOCO	GAN	Bi-LSTM
[117]	MultiMNIST, CLEVR, MSCOCO	GAN	char-CNN-RNN
[402]	COCO-Stuff, VisG	GAN	Conv-LSTM
[300]	Abstract Scenes, MSCOCO	Conv-RM	BiGRU
[395]	CUB, MSCOCO, Oxford-102	GAN	CNN, LSTM
[360]	MSCOCO, CUB	GAN	BiLSTM

3.8. Visual Retrieval

Most image retrieval works focus on fetching relevant images for a given textual query, represented by a few specific keywords describing attributes instead of the elongated textual descriptions. Such approaches have been widely used for retrieving products with similar concepts in fashion markets [107]. Other applications involve image tagging, text-to-image, and image-to-text retrieval tasks [296] based on accumulated concepts in the visual and semantic arena. Dong et al. [73] approached the retrieval task as a cross-media matching where either images are represented in the textual space or text is translated to appropriate visual embeddings to further use matching for relevant retrieval. Similar to this, several works pose the retrieval problem as a bidirectional task for sentence-to-image retrieval and vice-versa. In order to achieve this, more often than not, it is crucial to learn a shared embedding space for text and visual attributes for obtaining latent features before retrieval [82, 94, 120].

Many different variations have been proposed to the task specifics and underlying approaches for retrieval based objectives. Nagarajan and Grauman [222] separated objects and attributes while learning latent embeddings, thereby, ensuring that new attribute-object combinations when encountered, can be easily understood, instead of being mixed up. For retrieving similar yet specifically different images from the database, Vo et al. [330] inputted an image with a text-query describing necessary changes to be considered from the present image while searching for other relevant images for retrieval.

Apart from image retrieval, comes another analog where retrieval is based on more interactive queries as per user interaction [310]. Guo et al. [104] brought forth the novel task of dialogue-based retrieval where retrieval searches are based on agent-user interactions, which aided the establishment of user feedback in loop while retrieving relevant items from the database.

Table 8
Latest research in Visual Retrieval (VR)

Ref.	Dataset	Visual Encoder	Language Model
[107]	Fashion200k	CNN	BOW, word2vec
[76]	MS-COCO	CNN	BiLSTM
[172]	MS-COCO	Faster R-CNN	BERT
[181]	MS-COCO	Faster R-CNN	BERT
[346]	MS-COCO, Flickr30k	Faster R-CNN, ResNet-101	BiGRU

Trends in VR. Recent works in text-to-visual retrieval tasks have emphasized the learning of coherent VisLang representation spaces to obtain precise, meaningful matches (see Table 8). Two major trends that have recently spanned this entire research area mainly focus on unbiased extraction and feedback based. As user-feedback has been widely used in product searches [107], it has played a vital role in improving the performance in a loop-wise manner. More recently, deep learning frameworks have shifted the focus more from classic ranking or matching algorithms to discovering semantic concepts and cues in both textual and visual spaces [107, 157], either independently or combined.

4. Latest Trends in VisLang Modeling

In this section, we look at the latest papers in the multimodal application of VisLang research and observe the key modeling trends adopted by the papers.

4.1. Multimodal Representation Learning

Multimodal inputs including a visual input (image/video) and a textual input are either encoded individually to generate separate representations that are later fused or processed simultaneously using a network that directly generates a hybrid multimodal representation. Here, we focus on a diverse set of methods with shared multimodal latent space.

Visual Encoders. Visual encoders perform the task of extracting semantic information about key entities present in the visual inputs. They encode the input to a lower-dimensional manifold that captures dominant attributes and forms associations between them. This task of concealing complex visual inputs onto a denser feature space to perform a diverse range of downstream tasks is an age-old computer vision technique [318]. Multimodal approaches that form detached embeddings for visual inputs have often utilized popular image classification-based deep networks like LeNet [169], VGGNet [284], ResNet [112], or sometimes even a simple CNN to extract meaningful features from the input. Several approaches that require identifying key objects in visual input also employ prevailing deep object detection networks for generating embeddings that are later processed by further architecture. These object detection networks like RCNN [93], FasterRCNN [264], YOLO [261], etc. return the bounding boxes and the class predictions of the located objects present in the input, which are later used to correlate them with similar entities present in the language input as well.

Language Encoders. Frameworks that generate separate embeddings for each modality often consist of a temporal model that captures contextual relationships from text using vocabulary comprehension capabilities. As discussed in Devlin et al. [68], pre-trained language encoders can broadly be classified into two general categories, namely contextual and context-free. Context-free representation models like word2vec [213] and GloVe [240] generate embeddings for each word irrespective of its usage and surrounding words. On the other side, contextual representation models encode each word based on their contextual position in a given text. Further, we can divide contextual models into unidirectional and bidirectional. While unidirectional models comprehend the context of each word from one direction, bidirectional models examine the context from eithr side. Bidirectional Transformer networks like [68, 194] pre-trains an encoder to predict certain masked words, while learning to differentiate between positively and negatively correlated samples parallelly. Many multimodal systems have also employed simple temporal models like LSTMs, RNNs, or GRUs for generating text-based encodings.

Hybrid Representations. While prominent approaches in recent literature extracted vision and language features before fusing them, some approaches have also tried to directly embed a combined multimodal embedding from inputs. These methods obtain their motivation from lapsed classical Deep Boltzmann Machine (DBM) based multimodal generative models [290] directly processing the data to generate embeddings that could be deployed for a diverse set of classification and retrieval tasks. Later extended for temporal models, Rajagopalan et al. [258] proposed using multi-view LSTM for modeling view-specific and cross-view interactions over time to generate robust latent codes for image captioning and multimodal behavior recognition by directly undertaking the multimodal inputs. While the development of combined latent space by direct processing of the input modalities began the trend of generating joint embeddings that could be used for various tasks, the trend of generating individual spaces followed by fusion to obtain a generalized encoding has taken over for most VisLang tasks.

Multimodal Fusion. Multimodal fusion [13] is the amalgamation of individual embedding spaces corresponding to the visual and textual input to obtain a composite space that possesses knowledge of both: the semantic visual features as well as contextual language information, required for various VisLang tasks.

Hierarchical fusion that integrates two modalities at a time in the first step, followed by homogenization of all three modalities of text, audio, and visual inputs, has proven to be instrumental in tasks like sentiment analysis [205]. Fusion of these modalities has also been practiced via the virtue of low-range tensors [195] or greedy layer-sharing [125]. Besides, the task of multimodal fusion has also been posed as a neural architecture search algorithm over a space spanning assorted set of architectures [242]. While many prior approaches have reaped the benefits of bi-linear and tri-linear pooling for the combination of multimodal features, recent approaches have also utilized multi-linear fusion to incorporate higher-order interactions without any restrictions [124].

Multimodal learning is prone to a variety of challenges. Identified in Wang et al. [337], the prime sources of challenges are in overfitting due to sizeable architectures and contrasting learning rates of each modality. To cater to these problems, the authors proposed an optimal incorporation of each modality based on their overfitting trends. Other contemporary approaches have also dispensed the model with the freedom to decide the method to combine multimodal features, instead of fixing it apriori. Sahu and Vechtomova [266] proposed a network that progressively learns to encode significant features to model the context from each modality specific to the set of data provided.

Multi-view sequence learning is another VisLang avenue where fusion plays a crucial role. Zadeh et al. [387] introduced a *memory fusion network* that adjudges both view-specific and cross-view interactions to model time-varying characteristics.

4.2. Attention Mechanisms

4.2.1. Onset of Attention Mechanisms

The advent of deep learning era brought about an influx of works focussing on developing Seq2Seq models [299] aimed at generating meaningful output sequences based on an input sequence, both of arbitrary lengths. The initial works consisted of a language encoder and decoder, predominantly an LSTM/GRU, for tasks like machine translation [299] and video captioning [325].

One major drawback of such Seq2Seq models emerged out as their inability to accommodate long sentences [55]. The fixed-size context vectors failed to encapsulate information from longer sentences. As a result, these models often suffer from a sharp performance dip when processing complex language inputs. In order to cater to this problem, the first attention mechanism [14, 199] was introduced for the purpose of Neural Machine Translation (NMT).

Attention mechanisms in deep learning can be simply defined as channeling the importance of input regions based on certain factors and weighing them as per their influence.

4.2.2. Attention in VisLang

Soon after, followed the instigation of attention mechanism for other tasks like image captioning to anchor the objects of interest in visual inputs [359]. Here, we list the broad categories of attention utilized for a diverse range of VisLang tasks.

Soft and Hard Attention. In order to weigh the image regions based on their importance for a particular input, attention has been applied to the visual features extracted from images using a simple CNN encoder. Two broad categories of attention have been proposed to channelize the emphasis of different regions [359], namely the *deterministic soft* and *stochastic hard attention*. In *soft attention*, the attention map is multiplied to the extracted features and summed up to obtain the relevance of all image regions. In contrast, *hard attention* samples certain features based on a probability distribution to obtain the most relevant image region. In practice, *soft attention* is more popularly utilized because of the ease of applying gradient descent due to its differentiability. Most commonly, such attention methods have been applied to the visual model in the architecture.

Global and Local Attention. Initially introduced for NMT [199] this attention mechanism has been pivotal in developing a divergent set of VisLang tasks. This mechanism works on the principle of constructing a context vector as the weighted sum of hidden states of the temporal model, weights of which are learned by a separate alignment model. It enables the model to learn richer representations guiding it to pay attention to the more important input samples. In *global attention*, each of the states prior to the current state is taken into account while computing the output contrary to the *local attention*, where only a few states are utilized for the same. Predominantly, *global* and *local attention* is utilized in the language sub-network in VisLang models.

Self-Attention. Attention could also be applied within individual sequences to capture temporal relationships between components in order to generate better representations. While former approaches apply attention within input and output sequences, self-attention applies it within the input sequence itself in the encoding stage to generate better representations. Originated for the task of machine reading [50], it has spanned a diverse range of applications that include visual captioning [185], visual question answering [180], image-text matching [353], and many more.

A variety of works have utilized variants of self-attention for extricating semantic context. *Hard self-attention* has been extensively studied under the lens of the medical imaging, some of which focuses on medical image segmentation [228, 287], disease classification [100], *etc.* Hu et al. [126] used *soft self-attention* to reweight the channel-wise responses at a certain layer of a CNN to incorporate global information when making a decision. Such a network is very flexible due to the use of *soft attention* rather than a *hard* one, and can be combined with a wide range of architectures with ease.

Ramachandran et al. [259] proposed self-attention to be applied as a separate independent layer instead of the former application of serving as a simple augmentation on top of convolutional layers. Such an approach was proven to enhance image classification while using fewer parameters.

Despite its diversified applications and success in a wide range of vision and language tasks, this type of attention is less prevalent due to its high computational requirements in terms of both time and space.

4.2.3. Paradigm Shift in VisLang Tasks

The past decade has seen a drastic evolution of vision and language tasks that have transformed from simple tasks requiring processing of fused multimodal embeddings, to complex tasks that require higher-order reasoning and deep understanding of semantic contexts presented in the inputs. New tasks like VCR, VLN, MAC, etc. demand the model to not only comprehend natural language and identify objects in the scene, but also capture inherent relationships between individual entities present in the input. The model's ability to reason their predictions has become exceedingly essential, leading to the emergence of a new domain of interest referred to as *Explainable AI* [268]. Such systems not only generate explicable predictions but also are able to detect and therefore eradicate biases in the model that arise due to the ones present in training data.

Owing to these complex set of emerging tasks and the diverse set of datasets that have surfaced, a number of novel attention mechanisms have been utilized to embed the model with the deep contextual understanding.

Graph-based Attention. With the aim to perform reasoning about each entity, graph-based attention mechanisms have depicted an incredible ability to inculcate deep semantic relationships between independent entities extracted from the visual and language-based encodings.

Choi et al. [56] proposed using a directed acyclic graph-based attention for extracting domain knowledge to learn high-quality representation for healthcare applications. Later, a factor-graph based attention capable of combining any number of data utility representations was proposed for the task of visual dialogue [271].

It is often observed that graph-based attention mechanisms flourish in tasks that require user-specific feature extraction. Such frameworks tend to capture intrinsic relationships between encoded representations and features present in the data. Chen et al. [45] utilized graph-based attention for image captioning to add more control over how much fine-grained details are required in the caption by the user. Their graph composed of three types of nodes representing objects, attributes, and relationships based on which captions are generated. Also, a graph attention framework with multi-view memory was used for the task of top *n*-recommendations as per user-specific attributes [328].

Hierarchical Attention. In order to extract robust and meaningful semantic information from each individual element of text, hierarchical attention mechanism, introduced by Yang et al. [372], utilizes separate sentence and word encoders. This framework involves building separate sentence and documents embeddings using word and sentence hierarchy respectively by passing the output of the lower hierarchy on to the higher one.

Although, initially studied predominantly for language-based tasks like text summarization [69], document ranking for q&a [415], contextual image recommendation [351], hierarchical attention has also found immense amount of application in vision-based tasks like medical image segmentation [70], action recognition in videos [345], image captioning [336], video caption generation [288], crowd counting in images [285] as well.

Co-Attention. Attention may also be applied in a pairwise manner in order to learn affinity scores between two pieces of documents or texts mostly for matching-based applications. Such frameworks are most common in tasks that require comparison between text samples like essay scoring [394], text matching [308], reading comprehension [417]. Xiong et al. [357] utilized co-attentions for the fusion of independent question and document encodings for the purpose of question answering. In order to amalgamate the information retrieved from multimodal inputs (*i.e.* audio, text and video), Kumar et al. [164] performed multimodal alignment with the help of co-attention for multimodal question-answering. Li et al. [177] matched images to textual description by using two separate co-attention modules for extracting spatial and semantic information respectively.

This attention mechanism is one of the most profoundly utilized type for the task of VQA due to their ability to model corresponding words in the questions to objects in the visual input. Yu et al. [384] proposed self attention between images and questions alongside question-guided-attention for images as a separate layer in order to map key objects in questions to the ones detected in the images. Lu et al. [198] introduced a novel co-attention mechanism for jointly reasoning about the image and the question besides reasoning about the individual inputs in a hierarchical fashion. For video question answering, Li et al. [180] utilized co-attention for computing the important words in the questions besides self attention for computing the video features corresponding with respect to the input question. Lu et al. [198] used co-attention to jointly reason about images and questions in a hierarchical fashion for VQA. The ability of co-attention in building robust image-question representations has been illustrated in various works that include [385, 392, 223].

Others. A variety of other attention mechanitsms were also introduced with the aim to supplement visual language tasks with a reasoning-based backbone, making their predictions interpretable as well as effective. Yu et al. [383] extended the traditional self attention for unimodal inputs that captures inter-modal interactions to a unified attention framework that captures both inter as well as intra-modal interactions of multimodal features. They introduce a network that performs multimodal reasoning using gated self attention blocks for the tasks of VQA and visual grounding. Gregor et al. [98] utilized the selective attention mechanism introduced for the purpose of handwriting synthesis [96] and Neural Turing Machines [97], in order to generate high fidelity complex images indistinguishable to the human eye. Another work Pan et al. [232] focused on capitalizing on bilinear pooling blocks in order to discerningly attend to certain visual regions or performing multimodal reasoning. These *X-Linear Attention* blocks capture higher order feature interactions by utilizing spatial and channel-wise bilinear attention, leveraging them for the task of image captioning.

4.3. Transformers in Cross-Modal Research

4.3.1. Onset of Transformers for Capturing Temporal Data Characteristics

Transformers are architectures that take advantage of two separate networks, namely encoder and decoder, to transform one sequence into another. Unlike formerly described Seq2Seq models, transformers do not consist of a temporal model like an LSTM or a GRU. Initially, such models were found to be effective in generating sequences from the same distributions for tasks like machine translation [322]. Transformers are models employing attention mechanisms (described in Section 4.2) that capture temporal characteristics (predominantly in natural language processing). These typically undergo faster computation as they do not require sequential processing of data, therefore, promotes parallelization of data as opposed to RNN or LSTM-based temporal models. For the purpose of handling sizeable datasets, transformers outperforms its counterparts and hence, it has been active area of research in the VisLang community. After the advent of the popularity of simpler temporal models for the encapsulation of time-varying signals, various models arose that sought to replace the LSTM/RNN-based approaches as they demanded heavy computation requirements and often suffered from overfitting and vanishing gradients. Uncomplicated models like Temporal Convolutional Network (TCN) [15] and Gaussian-process VAEs [19] besides transformers have effectively replaced LSTMs and RNNs due to their ease of implementation, rapid and stable training, and generalization capacity.

4.3.2. Pre-Training Trends using Transformers

Recent literature has seen a sudden outburst of research interest in transformers for learning representations that can be utilized for a wide variety of tasks. Tan and Bansal [304] proposed LXMERT, a cross-modal transformer for encapsulating vision-language connections by utilizing three specialized encoders corresponding to object relationships, language and cross-modality to pretrain on five diverse tasks. Cho et al. [54] goes further with this idea to enable the model to generate images from these transformer representations via significant refinements in the training strategy, empowering the model to rival state-of-the-art generative models. Devlin et al. [68] introduced BERT that learned deep bidirectional representations from textual data to design a pre-trained model that can be fine-tuned for specific tasks like question answering and natural language inference.

Later, this framework was extended for multiple modalities to generate representations that encapsulate fused information from different sources of data like speech, text, and visual inputs (image or video) in a self-supervised fashion. VideoBERT Sun et al. [297] built upon this idea for creating such representations for videos that could be later used for downstream tasks like video captioning and action recognition. On similar lines, ViLBERT [196] and VisualBERT [176] generated shared representations for images and text enhancing performance on tasks like image captioning, visual question answering, commonsense reasoning, and image retrieval. ImageBERT [252] utilized weak supervision for generating an image-text joint space for unique and specific prediction tasks that involved input masking and text matching. More recently, contrastive learning paradigms have accentuated the competency of self-supervised learning using data augmentations in vision-based applications like image classification. When combined with multimodal representation learning using BERT-inspired frameworks, its efficacy in developing better representations has been exploited for tasks that involve joint training over multiple modalities [314, 295].

Such representations have been utilized for a diverse set of multimodal tasks that include visual question answering [409, 370, 5], visual captioning [409, 135], visual dialogue [221, 343], cross-modal retrieval [87], *etc.*

4.4. Evaluation Metrics

Metrics for language-based outputs popularly involve Bilingual Evaluation Understudy (*BLEU*) [233], Recall Oriented Understudy for Gisting Evaluation (*ROUGE*) [190], Metric for Evaluation of Translation with Explicit Ordering (*METEOR*) [17] and Consensus-based Image Description Evaluation (*CIDEr*) [324] across a variety of multimodal tasks. Originally introduced for machine translation tasks, *BLEU* score effectively evaluates any generated text compared to a reference text for a variety of tasks. It operates on counting matching *n*-grams, ranging from 0.0 (for a perfect mismatch) to a 1.0 (for a perfect match). *ROUGE* works on comparing a generated summary against target summary by considering the ratio of number of overlapping words and the total number of words. Some of its many variants are *ROUGE-1* (unigram overlap), *ROUGE-N* (N-gram overlap), *ROUGE-L* (based on Longest Common Subsequence). *METEOR* (originally introduced for machine translation tasks) is calculated via a weighted harmonic mean of unigram precision and recall, with a higher weight assigned to recall. To evaluate generated image descriptions based on human consensus, *CIDEr* measures sentence similarity of a generated sentence across a set of ground truth sentences considering factors like grammaticality, saliency, precision, and recall.

For the tasks with visual outputs, *R-precision* [61] was introduced for retrieval-based algorithms, later used for language-to-image generation task. It provides a ratio of *r* relevant retrievals given the top-R retrievals. Inception Score (*IS*), introduced by Salimans et al. [267] was to measure the quality of the generated samples in terms of semantically meaningful objects and diverse set of images, comparing marginal label distribution with conditional label distribution. *Fréchet Inception Distance* (*FID*) Heusel et al. [116] further improved upon IS by comparing generated samples against real samples instead of comparing with themselves. In contrast to *IS*, where higher scores are better, lower *FID* is better, denoting a lesser difference between the distributions of the generated and the real samples.

Besides these metrics, human evaluation via crowd-sourcing is another popular technique to assess the efficacy of predictions in VisLang tasks like VQA and IC.

Apart from these standard metrics and human evaluation, several recent works have proposed task-specific metrics. Here, we list the recent literature that have proposed novel evaluation metrics.

Metrics for IC. Various popular IC evaluation metrics are overly sensitive to n-gram overlap, as a result they do not correlate well with human assessment. To counter this, Anderson et al. [8] proposed *SPICE* to capture human judgment motivated by the importance of semantic proportional content over scene graphs. On similar lines, Sharif et al. [274] also introduced a learning-based metric that quantifies both the lexical and semantic correctness of the generated caption to improve correlation with human judgment. Despite the high association with human evaluation, these metrics fail to capture syntactical sentence structures. Therefore, Cui et al. [62] came up with an evaluation metric that is specifically modeled to distinguish between machine and human-generated visual captions. Likewise, Jiang et al. [141] proposed a novel metric called *TIGEr* that not only quantifies how well the caption captures the contents in the image but also their proximity to human-generated captions.

While captions containing similar words or their synonyms could be semantically dissimilar while ones not having any such similarities may be correlated semantically. In order to capture semantic similarities instead of commonalities in objects, attributes, or relations, Kilickaya et al. [152] evaluated the performance of the metric *Word Mover's Distance* [165] against other popular language metrics to compute the distance between documents.

Metrics for VQA. Various VQA datasets have skewed distributions due to the bias prevalent as a result of the varying number of samples present from each answer category. In order to cater to this problem of inductive data bias, Kafle and Kanan [143] proposed to utilize *Arithmetic and Harmonic mean-per-type (MPT)* of the accuracies obtained from each of the answer categories for a fairer evaluation. For specific subtasks of VQA wherein the questions are based on the texts found within the scene images, Biten et al. [22] introduced a novel metric *Average Normalized Levenshtein Similarity (ANLS)*, that quantifies OCR by keeping in mind the reasoning capability of the model and softly penalizing OCR mistakes. For the particular task of VQA when questions are expressed in two different languages, Wang et al. [340] proposed a metric called *Evidence-based Evaluation (EvE)*. This metric evaluates the model on two grounds namely correctness of the answer and sufficiency of evidence to support the predicted answer. Although, this work focused on a subtask of VQA, but this metric can be applied to general VQA settings as well, thereby making answer predictions from VQA models more justifiable.

Metrics for other VisLang Tasks. The task of visual question reasoning offers a challenging engagement in the sense that they seek to design models that possess high-order reasoning capabilities. In order to evaluate the capacity of the model to provide interpretable justifications, Cao et al. [31] suggested to utilize the *explainable evaluation metric* that calculates the triplet (the questions often contain relationship triplets to enable the model to perform multistep reasoning) precision for each question and average recall of all q&a pairs to obtain the final recall.

Hudson and Manning [134] proposed a novel dataset for visual reasoning and compositional question answering, accompanied by a set of original evaluation metrics to enumerate the performance of such models. The authors defined the following metrics for inference: *consistency* that utilized questions' semantic representation for inferring the associations between them, *validity and plausibility* that verifies if the answer lies within the scope of the question, *distribution* that validates if the approach is able to model the conditional distribution of the answers, and *grounding* that verifies the relevance of attended regions with respect to the questions.

5. Emerging Ideas in VisLang Research

In addition to the task-specific and modeling trends in recent VisLang literature, we also identify the emerging topics that leverage VisLang data using unique training strategies. The last few years have witnessed an escalation in these methodologies being applied to the VisLang domain.

5.1. Multi-task Learning

Multi-task Learning (MTL) [33] is an age-old concept of joint learning across multiple tasks to transfer the learnings of one task onto the other, eventually benefiting the performance on each of them. This approach of MTL has been gaining popularity in the domain of vision and language as a result of the implicit manifestation of similar attributes between different modalities. The encapsulation of attributes from multimodal data sources could help the model capture semantic relationships between entities of varying modalities, enhancing the model's understanding of concepts present in the data.

There have been attempts to create a model ViLBERT-MT [197] (MTL extension of ViLBERT [196]) capable of learning a joint representation for four diverse VisLang tasks on a collection of 12 datasets by employing a large-scale MTL training procedure followed by fine-tuning for specific tasks. Some works [224, 4] also focused on learning hierarchical representations wherein predictions for multiple tasks could be performed at different levels of hierarchy. lamBERT [217] extended the BERT framework for generating multimodal representations by blending it with reinforcement learning in order to perform MTL along with transfer learning.

MTL frameworks have also extracted meaningful representations for specific VisLang tasks by dividing them into sub-tasks, each of which combines to solve the complete objective. IC has one such task that has greatly benefited from the advent of MTL in deep learning systems. Some approaches [404] have tried breaking down the unabridged task of captioning into a set of three sub-tasks that include learning a category-aware representation, syntax generation model, and captioning model. Such models hypothesize that treating object classification and syntax knowledge as key aspects of IC and collectively applying MTL for optimizing these three objectives would lead to models with better cognition capabilities coupled with syntax understanding of natural language. On similar lines, IC has also been proposed as a task of learning captioning besides another supplemental task like activity recognition in visual inputs [79], image-sentence retrieval [332] or text-to-image synthesis [364]. While there have been attempts to disintegrate the task of IC into several sub-tasks to enhance the performance of IC models, some works [327] have also focused on using image captions as an auxiliary source of data for other tasks as it promotes the model to apprehend associations between entities present in visual and language inputs. Apart from the deep learning-based MTL approaches, some works [175] have also focused on designing a reinforcement learning approach that strives to construct several tasks (like rewards, attributes, captions) for captioned videos in order to strengthen the generalization power of their IC model, using much fewer computational resources.

Although MTL is more commonly applied for the task of IC, they have also been several attempts to apply it for the prime task of VQA. Here, the concept of MTL is embodied by breaking into sub-problems, where each contributes towards generating reasonable answers to questions given an image [245]. Also, Kornuta et al. [159] utilized MTL alongside transfer learning for capturing similarities between distributions of radiology images coming from different modalities to perform four disjoint question-answers tasks on them.

In order to bridge the gap between the performance of VLN models on previously trained and unseen environments, some works have focused on employing an MTL framework for learning an environment-agnostic latent representation that could be utilized for generalizing on unseen environments as well [339]. Deep reinforcement learning frameworks, amalgamated with a novel dual-attention have also been put to good use to disentangle features generated from textual and visual inputs [35]. These representations were then utilized for tasks like semantic goal navigation and embodied question answering.

Some approaches have also pursued performing a diverse set of unique multimodal tasks using an MTL framework. Learning sentiment analysis alongside emotion recognition by capitalizing on visual, textual, and acoustic data from video frames via a context-level attention mechanism achieved satisfactory results on both tasks [3].

5.2. Interpretability and Explainability

Fairness in machine learning has recently raised many questions about the transparency of algorithms that were otherwise used as black boxes. This has brought forth several biased assumptions made by models due to the inherent biases present in the data and prior knowledge occupied by the models, which previously remained unnoticed. Interpretability and explainability are two such aspects of this trend. On the one hand, interpretability requires understanding the cause and effect relationship of certain learnt factors and answering the "*what*" of the underlying mechanics [92], explainability focuses on explicitly describing the facts regarding "*why*" and "*how*" [355]. However, more often than not, these terms go hand in hand, directing to explicability of the system.

Several such efforts have been recently made in tasks such as VQA, mainly leveraging attention-based textual to visual inferences using parse-trees [32] and, qualitatively visualizing the effect of altering textual inputs keeping the image fixed [323] using a probabilistic approach. Interpretability has also been explored by analyzing the accuracy capability of user predictions on VQA agents in interactive learning settings [127, 7]. Modular approaches tend to have a higher degree of interpretability. The multimodal task of VQA becomes more interpretable and coherent with multimodal interpretability as well wherein the choice answers can be justified in both textual and visual arena [235]. Also, reasoning out the question-answering task has been well captured with learning question specific graph-based interactions in images [227] and hierarchical patterns to provide valid explanations and answer-specific substreams in sequential data using visual-text attention [187].

Other approaches move one step beyond visualizing intermediate effects of attention to studying more task-relevant attributes for reasoning out the behavior of deep learning models. For example, explainability in image captioning is achieved at pixel-level by showing the relevance of specific keywords in textual descriptions with relevant entities in the visual using layer-wise relevance backpropagation (*LRP*) [295], even in medical images [286]. More generally, multimodal explainability has gained visibility to provide coherent reasons in both textual and visual spaces for model predictions, leading to significantly vivid and self-explicable models [234].

Visual Reasoning task requires machines to ideally look beyond the face value of any image to capture correct relations and context before generating suitable descriptions. This has been benefited by using scene graphs as inductive biases [278], Raven's Progressive Matrices paired with structured graphical representations [393].

Some other prominent approaches for building interpretable and causal models is via disentangled representations, multimodal explanations and counterfactuals [92, 146].

Datasets. Several new datasets have been proposed for achieving interpretable multimodal learning. Zhang et al. [393] proposed a new dataset based on Raven's Progressive Matrices (RPM)[2] to facilitate the task of reasoning. It is curated to emphasize visual recognition reasoning, comprising images and related RPM problems, with tree-structured annotations. A counting-based dataset is sampled from the available *VQA 2.0* [2] and *Visual Genome* (VG) [163] datasets for the task-specific release by Trott et al. [315]. This work focused on countable quantitative question answering for answering specific queries asking "*how many?*". Park et al. [234] introduced two novel datasets dedicated to explainability for visual question answering *(VQA-X)* and activity recognition *(ACT-X)* tasks comprising of textual justifications for each image-text input pair. The *VQA-X* dataset has since then been considered a benchmark for many other explainable models [238, 348].

Contrastive Learning. Contrastive Learning provides neural models with self-supervised competence using relevant (positive) and irrelevant (negative) pairs. More recently, it has been utilized to improve multimodal representations, be it for pretraining [280], where it helps in diminishing issues of noisy labels and domain-biases or for enhancing task-specific performances. For the latter, it has been explored in different capacities for the task of image-captioning for either promoting distinctiveness in the generated captions [63] or mapping regions in the image with relevant words [105], in order to be leveraged for appropriate attention-weights. Additionally, contrastive loss has also been used to model inter-class dynamics in multimodal settings to enforce modality-agnostic feature representations with high semantic interpretability for multiple downstream tasks [319].

Counterfactuals. Another important aspect of interpretable machine learning models is explored under the lens of counterfactual reasoning. It aims at inferring the causes of a prediction and the relationship between them under distortions in the input. Most recent approaches tend to capture the effect of masking essential input objects (visual) [1] or tokens (textual) or both [41] and analyze how it deviates from the original image. Such methods allow more interpretable machine predictions with supporting cause-effect relations. VQA models are generally considered language biased. To capture the intricacies of this fact, Niu et al. [225] studied this causal inference using counterfactual settings where visual ground truth input is considered absent in an imagined scenario. This inference strategy explains inherent lingual biases present in VQA models. Similarly, for analyzing the effect of visual biases, Pan et al. [230] synthesized similar yet different images than ground truth and then studied how and why the answers change with differing visual distortions.

For visual captioning, *counterfactual explanations* help immensely in analyzing the learning patterns of the models and the reasons behind certain predictions. Such explanations emphasize the observations, present or missing, that lead to certain outputs [113, 146]. Fang et al. [78] obtained counterfactual resilience in image descriptions by parsing entities, semantic attributes, and color information separately. Moreover, such counterfactual reasonings have been utilized even in reinforcement learning settings for non-auto-regressive image captioning [102] and scene graph representation [42] to optimize team rewards as per individual agent counterfactual baselines in a multi-agent environment.

Bias and Fairness in VisLang. Lack of balanced data and feature selection have been commonly prone to introducing biases in models and machine learning algorithms. This often leads to a compromise on the fairness and transparency of such models on various grounds. Multimodal representations being subjected to more than one type of information, can infuse multiple such instances of biased information into deep learning models. Peña et al. [239] demonstrated how biases affect automated recruitment systems in one way or the other. With varying gender and ethnicity records across the dataset, the deep learning frameworks pick up subtle biased information even when certain information modalities are masked out from the input.

Bias trends have also been observed in task-specific trends.

[2] RPMs are reasoning-based questions comprising of visual geometric patterns with sequential non-verbal cues, with the missing piece to be identified.

In VQA task, the models often pick up statistical irregularities, thereby inducing biases in the model predictions and generations. Unimodal biases in the textual inputs neglect visual information, thereby reducing multimodality considerations. Such biases often lead to massive drops in the performance when confronted with data outside training distributions [25]. Moreover, most models show that generalized and trivial questions are commonly answered with prior lingual knowledge instead of querying the image. Therefore, keyword dependencies over correct image reasons are necessary to obtain correct, interpretable models and can be comprehended via attention maps [206]. Other adversarial and discriminative methods have helped to overcome language bias by analyzing question-only approaches where visual modality is masked to see the partial or complete influence of language statistical patterns [260].

For the task of image captioning, visual cues in the training images serve as potential bias carriers, which are further amplified in the model predictions during inference. Several such biases have been seen in identifying gender correctly. Most model predictions base unclear gender descriptions as per activity and context in the image, thereby adding unfair inductive biases, hampering gender-neutral understanding of models [23]. Other such efforts focus on two major subtasks or gender-neutral captioning in case of occlusions and correct gender classification otherwise [20].

All in all, such methods focus on making different multimodal frameworks more reliable, interpretable by allowing the models to provide reasonable predictions for the right reasons rather than just optimizing the performance without looking for deep-down causes of errors.

5.3. Domain Adaptation in VisLang

Domain adaptation is simply the procedure to learn a representation or model for the source domain and evaluating it on the target domain. Typically in initial unsupervised approaches [86], the labels for source domain were utilized for achieving generalization on the target domain with incomplete or no labels by deploying two separate classifiers for domain and label classification, respectively.

Learning domain generalizable representations for the task of VLN is indispensable due to the high cost of training in the real world. Commonly, several approaches aim to learn representations on simulations that would extrapolate in the real world scenario. Other works have focused on learning transferable representations that could enable training on one domain and later transferring them to the target domain for VLN tasks like Room-to-Room (R2R) [128].

Various approaches have also aimed at using domain adaptation for VQA and IC. It is essential to design models that are generalizable to a broad set of datasets. For domain adaptation in VQA, various approaches aim at converting feature representations from one distribution (dataset in this case) to some other target distribution without sufficient labels. While some methods achieve this by maximizing the likelihood of answering questions in the target domain [34], others capitalize on limited labels via fine-tuning after training on source domain [362]. For IC task, learning user-specific personalized image captions requires the model to captures similarities and correlations between a collection of data samples coming from a distribution. Chen et al. [47] targeted transferring the learnt IC model trained on the paired large scale source dataset to target dataset with no paired data. This work utilized two critic networks besides the image captioner, namely the domain critic and the multimodal critic. While the domain critic aimed at making the source captions indistinguishable from the ones in the target, the multimodal critic predicts if the given pair is valid.

Other multimodal tasks like sentiment analysis, multimodal retrieval, etc. have also benefited from the emergence of deep learning-based domain adaptation frameworks. Prior work [253] has focused on identifying correlations between embeddings from different modalities using a multimodal attention mechanism, fusing these attended features and learning domain-invariant features by involving certain domain constraints in the optimization objective [111].

5.4. Zero-Shot Learning

Learning to generalize at inference time on samples from classes unseen during the training phase, referred to as Zero-Shot Learning (ZSL), has been extensively investigated in the vision and language paradigm individually. In recent literature, various approaches have tried to counter the lack of labeled examples of a certain set of data by deploying ZSL-based methods.

To solve the problem of lack of generalization of VC models on unseen objects, Demirel et al. [66] utilized a ZSL-based object detector model for identifying key objects in visual input along with a sentence generator using extracted features to produce captions.

ZSL in VQA aims at developing intelligent agents that comprehend concepts learned from one module (*i.e.* questions) and are capable of transferring this knowledge onto other modules (*i.e.* answers) during test time. For this, Li et al. [184] proposed a *zero-shot transfer* VQA dataset that reorganized the VQA v2 dataset in a manner that the words are divided among the different modules in an exhaustive and disjoint fashion. Teney and Hengel [309] also promoted ZSL in VQA by suggesting effective strategies that included pre-trained word embeddings, object classifiers with semantic embeddings, and test-time retrieval of example images that enhanced the zero-shot performance of existing approaches.

5.5. Adversarial Attacks

Attacks on machine learning models intended by the user to cause a false prediction using carefully designed examples (known as adversarial examples) are called adversarial attacks [95]. Adversarial attacks for general machine learning models (predominantly classification models) have been an active area of research for decades now, with tons of papers focusing on various novel types of attacks and others focusing on building models' defense mechanisms. A model's response to such adversarial examples justifies the generalizability to samples not present in the training set but belonging to the same distribution. Recently, the paradigm of adversarial attacks and the development of response against them via adversarial training have gained traction for multimodal tasks involving vision and language.

Although recent VQA models have spotted significant progress in performance, adversarial examples may hinder their practical application [361]. To evaluate the robustness of state-of-the-art VQA models, Sharma et al. [276] came up with an attention-guided adversarial sample generation technique. They also proposed an additional evaluation metric that quantifies the strength of a given attack based on relative decrease in accuracy and noise induced. Other approaches [129] have also tried to utilize semantically related questions and dubbed basic questions as noise to evaluate the robustness of these models. After extensive experimentation for determining if VQA models apprehend the importance of various inputs, Mudrakarta et al. [220] concluded that these models often fail to capture important question terms. Motivated by this, they proposed an attack that perturbs the question terms to fool VQA models. Contrary to approaches manipulating one input to the model, some have also tried out modifying multiple modalities to reduce the accuracy of these models. Tang et al. [306] proposed to manipulate both image and question and subsequently trained a VQA model adversarially to defend against such attacks.

Adversarial examples have also been successful in reducing the performance of VC models by adding noise to visual inputs. Zhang et al. [399] proposed to craft an adversarial example with semantic embedding of target captions to fool image captioning CNN-RNN-based models. On similar lines, Chen et al. [39] also introduced an adversarial attack for similar models that test the ability of model to be misled to produce a certain randomly chosen caption (or keywords). Another similar evaluation protocol utilized by other approaches [363] was to test if it was possible to generate a certain partial (words are some locations are restricted while other locations are not) caption using some perturbation in the input image.

6. Discussion

VisLang learning involves effectively transferring knowledge across modality spaces and uniting bi-modal representations in a collaborative structure. Such learning makes it essential to have robust representations for a generalized improvement in the performance of underlying downstream tasks. The rapid boom in VisLang research has accelerated the instigation of self-reliant models that interpret interactions in visual and language modalities. Despite the furtherances, there lies a plethora of challenges and future directions in this active research area.

Challenges. With the challenge of lack of available labeled data, along with unsupervised (or weakly supervised) approaches, unsupervised metrics are essential for fairly evaluating progressive approaches, in ways close to human evaluation. Several foregoing approaches that claim unsupervised nature of their learning methodology are not unsupervised in a true sense since they require labels at the evaluation stage. Therefore, standardized evaluation protocols for the inference of VisLang systems without the requirement of labels is indispensable. Hessel and Lee [115] proposed a novel diagnostic method for learning cross-modal interactions in multimodal learned representations. Moreover, video-based VisLang research tasks can add another overhead of temporal data to VisLang research, which differentiates it primarily from multimodal tasks. Such consideration shall bring forth yet another challenge of temporal alignment across modalities, which is generally latent in current vision-language studies. Another major challenge is the substantial efficiency in the alignment of representations across modalities, which has a significant impact on several downstream task performances. The lack of thorough multimodal inputs might still be an impending challenge, especially for temporal modalities such as videos, where missing or corrupted frames occur in practical scenarios. This calls for models where the processing of inputs is preceded by prediction of missing information, adding to the uncertainty of prediction.

Future Directions. Contrastive learning, probabilistic graphical models (like causal networks and counterfactuals), and disentanglement (popularly inspired by visual representation learning) have more recently paved a way into VisLang Research. It opens a wide arena of interpretable and transparent deep learning algorithms, thereby reducing overall bias and increasing the reliability of a system. Simultaneously, it opens the requirement for novel reasoning-based datasets for explicable models and relevant metrics that quantify not only the separation from ground-truth data but also measure the higher-level reasoning and cognition capabilities over complex datasets.

Heading towards more challenging and real-world intelligent systems, it is essential to step towards complex forms of current problems with minimum assumptions and inductive biases. This can be in the form of subjective question-answering, dialogue-based agents for caption generation, or generalizing vision-language navigation to multi-agent systems and unseen environments.

Emphasizing the generalization capability of algorithms to be deployed in real-world scenarios, multi-task learning, transfer learning, curriculum learning, reinforcement learning, zero-shot learning (ZSL), and unsupervised/self-supervised methods open a yet to be exhausted avenue of research for many VisLang tasks. Often regarded as the extreme case of domain adaptation, ZSL has been instrumental in the development of generalizable VisLang models.

7. Conclusion

We present and categorize the current VisLang tasks based on their key characteristics highlighting their prominent similarities and dissimilarities. The brisk developments in deep learning architectures have enabled the circumstance of compelling VisLang models that have outperformed humans in a diverse set of tasks. We outline the diversified applications involving vision and language modalities to design intelligent VisLang models with interpretable semantic cognition capabilities coupled with a comprehensive understanding of natural language. Further, we enlist the recent trends within each task and the learning methodologies harnessed in contemporary literature.

Acknowledgement

This research is supported by A*STAR under its RIE 2020 Advanced Manufacturing and Engineering (AME) programmatic grant, Award No. - A19E2b0098, Project name - K-EMERGE: Knowledge Extraction, Modelling, and Explainable Reasoning for General Expertise. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of A*STAR.

References

[1] Agarwal, V., Shetty, R., Fritz, M., 2019. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. ArXiv abs/1912.07538.

[2] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D., 2015. Vqa: Visual question answering. IJCV 123, 4–31.

[3] Akhtar, M.S., Chauhan, D.S., Ghosal, D., Poria, S., Ekbal, A., Bhattacharyya, P., 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. ArXiv abs/1905.05812.

[4] Al-Rawi, M., Valveny, E., 2019. Compact and efficient multitask learning in vision, language and speech, in: ICCV Workshop, pp. 2933–2942.

[5] Alberti, C., Ling, J., Collins, M., Reitter, D., 2019. Fusion of detected objects in text for visual question answering, in: EMNLP/IJCNLP.

[6] Alikhani, M., Sharma, P., Li, S.J., Soricut, R., Stone, M.B., 2020. Clue: Cross-modal coherence modeling for caption generation. ArXiv abs/2005.00908.

[7] Alipour, K., Schulze, J.P., Yao, Y., Ziskind, A., Burachas, G., 2020. A study on multimodal and interactive explanations for visual question answering, in: SafeAI@AAAI.

[8] Anderson, P., Fernando, B., Johnson, M., Gould, S., 2016. Spice: Semantic propositional image caption evaluation, in: ECCV.

[9] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018a. Bottom-up and top-down attention for image captioning and visual question answering. CVPR , 6077–6086.

[10] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I.D., Gould, S., van den Hengel, A., 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR , 3674–3683.

[11] Andreas, J., Klein, D., 2015. Alignment-based compositional semantics for instruction following, in: EMNLP.

[12] Aneja, J., Agrawal, H., Batra, D., Schwing, A.G., 2019. Sequential latent spaces for modeling the intention during diverse image captioning. ICCV , 4260–4269.

[13] Atrey, P., Hossain, M., El Saddik, A., Kankanhalli, M., 2010. Multimodal fusion for multimedia analysis: A survey. Multimedia Syst. 16, 345–379.

[14] Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473.

[15] Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. ArXiv abs/1803.01271.

[16] Baltrušaitis, T., Ahuja, C., Morency, L., 2019. Multimodal machine learning: A survey and taxonomy. IEEE TPAMI .

[17] Banerjee, S., Lavie, A., 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: IEEvaluation@ACL.

[18] Barezi, E.J., Fung, P., . Modality-based factorization for multimodal fusion, in: RepL4NLP@ACL 2019, pp. 260–269.

[19] Bhagat, S., Uppal, S., Yin, V.T., Lim, N., 2020. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. ECCV .

[20] Bhargava, S., Forsyth, D., 2019. Exposing and correcting the gender bias in image captioning datasets and models. ArXiv abs/1912.00578.

[21] Bhattacharya, N., Li, Q., Gurari, D., 2019. Why does a visual question have different answers? ICCV , 4270–4279.

[22] Biten, A.F., Tito, R., Mafla, A., Gómez, L., Rusiñol, M., Valveny, E., Jawahar, C.V., Karatzas, D., 2019. Scene text visual question answering. ICCV , 4290–4300.

[23] Burns, K., Hendricks, L.A., Darrell, T., Rohrbach, A., 2018. Women also snowboard: Overcoming bias in captioning models, in: ECCV.

[24] Cadène, R., Ben-younes, H., Cord, M., Thome, N., 2019a. Murel: Multimodal relational reasoning for visual question answering. CVPR , 1989–1998.

[25] Cadène, R., Dancette, C., Ben-younes, H., Cord, M., Parikh, D., 2019b. Rubi: Reducing unimodal biases in visual question answering, in: NeurIPS.

[26] Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., van de Weijer, J., 2016a. Does multimodality help human and machine for translation and image captioning?, in: WMT.

[27] Caglayan, O., Barrault, L., Bougares, F., 2016b. Multimodal attention for neural machine translation. ArXiv abs/1609.03976.

[28] Cai, G., Xia, B., 2015. Convolutional neural networks for multimedia sentiment analysis, in: NLPCC.

[29] Calixto, I., Liu, Q., 2017. Incorporating global visual features into attention-based neural machine translation., in: EMNLP, pp. 992–1003.

[30] Calixto, I., Rios, M., Aziz, W., 2019. Latent variable model for multi-modal translation, in: ACL, pp. 6392–6405.

[31] Cao, Q., Li, B., Liang, X., Lin, L., 2019a. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. ArXiv abs/1909.10128.

[32] Cao, Q., Liang, X., Li, B., Lin, L., 2019b. Interpretable visual question answering by reasoning on dependency trees. IEEE TPAMI .

[33] Caruana, R., 1997. Multitask learning. Machine Learning 28.

[34] Chao, W.L., Hu, H., Sha, F., 2018. Cross-dataset adaptation for visual question answering. CVPR , 5716–5725.

[35] Chaplot, D.S., Lee, L., Salakhutdinov, R., Parikh, D., Batra, D., 2019. Embodied multimodal multitask learning. ArXiv abs/1902.01385.

[36] Chaudhry, R., Shekhar, S., Gupta, U., Maneriker, P., Bansal, P., Joshi, A., 2020. Leaf-qa: Locate, encode & attend for figure question answering, in: WACV.

[37] Chen, F., Ji, R., Ji, J., Sun, X., Zhang, B., Ge, X., Wu, Y., Huang, F., Wang, Y., 2019a. Variational structured semantic inference for diverse image captioning, in: NeurIPS, pp. 1931–1941.

[38] Chen, H., Li, J., Hu, X., 2020a. Delving deeper into the decoder for video captioning. ArXiv abs/2001.05614.

[39] Chen, H., Zhang, H., Chen, P.Y., Yi, J., Hsieh, C.J., 2018a. Attacking visual language grounding with adversarial examples: A case study on neural image captioning, in: ACL.

[40] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y., 2020b. Counterfactual samples synthesizing for robust visual question answering, in: CVPR.

[41] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y., 2020c. Counterfactual samples synthesizing for robust visual question answering. ArXiv abs/2003.06576.

[42] Chen, L., Zhang, H., Xiao, J., He, X., Pu, S., Chang, S.F., 2019b. Counterfactual critic multi-agent training for scene graph generation. ICCV , 4612–4622.

[43] Chen, M., Kang, S.G., 2017. (re) live photos : Generating videos with gans.

[44] Chen, S., Jin, Q., Fu, J., 2019c. From words to sentences: A progressive learning approach for zero-resource machine translation with

visual pivots, in: IJCAI.

[45] Chen, S., Jin, Q., Wang, P., Wu, Q., 2020d. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. ArXiv abs/2003.00387.

[46] Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., Luo, J., 2018b. "factual" or "emotional": Stylized image captioning with adaptive learning and attention, in: ECCV.

[47] 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. ICCV , 521–530.

[48] Chen, X., Zitnick, C.L., 2015. Mind's eye: A recurrent visual representation for image caption generation. CVPR , 2422–2431.

[49] Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., jing Liu, J., 2019d. Uniter: Universal image-text representation learning. CVPR .

[50] Cheng, J., Dong, L., Lapata, M., 2016. Long short-term memory-networks for machine reading. ArXiv abs/1601.06733.

[51] Cheng, Y., Gan, Z., Li, Y., Liu, J., Gao, J., 2018. Sequential attention gan for interactive image editing via dialogue. ArXiv abs/1812.08352.

[52] Chetty, G., Wagner, M., 2008. A multilevel fusion approach for audiovisual emotion recognition, in: AVSP.

[53] Chi, T.C., Eric, M., Kim, S., Shen, M., Hakkani-Tür, D.Z., 2020. Just ask: An interactive learning framework for vision and language navigation, in: AAAI.

[54] Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., Kembhavi, A., 2020. X-lxmert: Paint, caption and answer questions with multi-modal transformers. ArXiv abs/2009.11278.

[55] Cho, K., van Merrienboer, B., Çaglar Gülçehre, Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation, in: EMNLP.

[56] Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J., 2017. Gram: Graph-based attention model for healthcare representation learning. ACM SIGKDD .

[57] Chou, S.H., Chao, W.L., Lai, W.S., Sun, M., Yang, M.H., 2020a. Visual question answering on 360deg images, in: WACV.

[58] Chou, S.H., Chao, W.L., Sun, M., Yang, M.H., 2020b. Visual question answering on 360deg images. WACV , 1596–1605.

[59] Cornia, M., Baraldi, L., Cucchiara, R., 2019a. Show, control and tell: A framework for generating controllable and grounded captions. CVPR , 8299–8308.

[60] Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R., 2019b. M2: Meshed-memory transformer for image captioning. ArXiv abs/1912.08226.

[61] Craswell, N., 2009. R-Precision. Springer US. pp. 2453–2453.

[62] Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S.J., 2018. Learning to evaluate image captioning. CVPR , 5804–5812.

[63] Dai, B., Lin, D., 2017. Contrastive learning for image captioning, in: NIPS.

[64] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M.F., Parikh, D., Batra, D., 2017a. Visual dialog. CVPR , 1080–1089.

[65] Das, A., Kottur, S., Moura, J.M.F., Lee, S., Batra, D., 2017b. Learning cooperative visual dialog agents with deep reinforcement learning, in: ICCV.

[66] Demirel, B., Cinbis, R.G., Ikizler-Cinbis, N., 2019. Image captioning with unseen objects, in: BMVC.

[67] Deshpande, A., Aneja, J., Wang, L., Schwing, A.G., Forsyth, D.A., 2019. Fast, diverse and accurate image captioning guided by part-of-speech. CVPR , 10687–10696.

[68] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: in NAACL-HLT.

[69] Diao, Y., Lin, H., Yang, L., chao Fan, X., Chu, Y., Wu, D., Zhang, D., Xu, K., 2020. Crhasum: extractive text summarization with contextualized-representation hierarchical-attention summarization network. Neural Computing and Applications , 1–13.

[70] Ding, F., Yang, G., Liu, J., Wu, J., Ding, D., Xu, J., Cheng, G., Li, X., 2019. Hierarchical attention networks for medical image segmentation. ArXiv abs/1911.08777.

[71] Do, T., Do, T.T., Tran, H., Tjiputra, E., Tran, Q.D., 2019. Compact trilinear interaction for visual question answering, in: ICCV.

[72] Dognin, P.L., Melnyk, I., Mroueh, Y., Ross, J., Sercu, T., 2018. Improved image captioning with adversarial semantic alignment. ArXiv abs/1805.00063.

[73] Dong, J., Li, X., Xu, D., 2018. Cross-media similarity evaluation for web image retrieval in the wild. IEEE Transactions on Multimedia 20, 2371–2384.

[74] El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L.E., Kahou, S.E., Bengio, Y., W.Taylor, G., 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. ICCV , 10303–10311.

[75] Elliott, D., Frank, S., Sima'an, K., Specia, L., 2016. Multi30k: Multilingual english-german image descriptions. ArXiv abs/1605.00459.

[76] Engilberge, M., Chevallier, L., Pérez, P., Cord, M., 2019. Sodeep: A sorting deep net to learn ranking loss surrogates. CVPR , 10784–10793.

[77] Fan, Z., Wei, Z., Wang, S., Huang, X., 2019. Bridging by word: Image grounded vocabulary construction for visual captioning, in: ACL.

[78] Fang, Z., Kong, S., Fowlkes, C.C., Yang, Y., 2019. Modularized textual grounding for counterfactual resilience. CVPR , 6371–6381.

[79] Fariha, A., 2016. Automatic image captioning using multitask learning, in: NIPS, pp. 11–20.

[80] Feng, Y., Ma, L., Liu, W., Luo, J., 2019. Unsupervised image captioning. CVPR , 4120–4129.

[81] Ferraro, F., Mostafazadeh, N., Huang, T.H., Vanderwende, L., Devlin, J., Galley, M., Mitchell, M., . A survey of current datasets for vision and language research, in: EMNLP, 2015.

[82] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. Devise: A deep visual-semantic embedding model, in: NIPS.

[83] Fu, T.J., Wang, X., Peterson, M., Grafton, S.T., Eckstein, M., Wang, W.Y., 2019. Counterfactual vision-and-language navigation via adversarial path sampling. ArXiv abs/1911.07308.

[84] 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 abs/1606.01847.

[85] Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., jing Liu, J., 2020. Large-scale adversarial training for vision-and-language representation learning. ArXiv abs/2006.06195.

[86] Ganin, Y., Lempitsky, V.S., 2015. Unsupervised domain adaptation by backpropagation, in: ICML.

[87] Gao, D., Jin, L., Chen, B., Qiu, M., Wei, Y., Hu, Y., Wang, H.M., 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. ArXiv abs/2005.09801.

[88] Gao, J., Wang, S., Wang, S., Ma, S., Gao, W., 2019a. Self-critical n-step training for image captioning. CVPR , 6293–6301.

[89] Gao, P., You, H., Zhang, Z., Wang, X., Li, H., 2019b. Multi-modality latent interaction network for visual question answering, in: ICCV.

[90] Garcia, N., Otani, M., Chu, C., Nakashima, Y., 2020. Knowit vqa: Answering knowledge-based questions about videos, in: AAAI.

[91] Ge, H., Yan, Z., Zhang, K., Zhao, M., Sun, L., 2019. Exploring overall contextual information for image captioning in human-like cognitive style. ICCV , 1754–1763.

[92] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning. DSAA , 80–89.

[93] Girshick, R.B., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR , 580–587.

[94] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S., 2014. Improving image-sentence embeddings using large weakly annotated photo collections, in: ECCV.

[95] Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. ArXiv .

[96] Graves, A., 2013. Generating sequences with recurrent neural net-

works. ArXiv abs/1308.0850.

[97] Graves, A., Wayne, G., Danihelka, I., 2014. Neural turing machines. ArXiv abs/1410.5401.

[98] Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D., 2015. Draw: A recurrent neural network for image generation. ArXiv abs/1502.04623.

[99] Gu, J., Joty, S.R., Cai, J., Zhao, H., Yang, X., Wang, G., 2019. Unpaired image captioning via scene graph alignments. ICCV , 10322–10331.

[100] Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y., 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. ArXiv abs/1801.09927.

[101] Guo, D., Xu, C., Tao, D., 2019. Image-question-answer synergistic network for visual dialog. CVPR , 10426–10435.

[102] Guo, L., Liu, J., Zhu, X., He, X., Jiang, J., Lu, H.Q., 2020a. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. ArXiv abs/2005.04690.

[103] Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.Q., 2020b. Normalized and geometry-aware self-attention network for image captioning. ArXiv abs/2003.08897.

[104] Guo, X., Wu, H., Cheng, Y., Rennie, S., Feris, R.S., 2018. Dialog-based interactive image retrieval, in: NeurIPS.

[105] Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D., 2020. Contrastive learning for weakly supervised phrase grounding. ArXiv abs/2006.09920.

[106] Han, C.W., Mao, J., Gan, C., Tenenbaum, J., jun Wu, J., 2019. Visual concept-metaconcept learning, in: NeurIPS.

[107] Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S., 2017. Automatic spatially-aware fashion concept discovery. ICCV , 1472–1480.

[108] Hao, W., Li, C., Li, X., Carin, L., Gao, J., 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. CVPR , 13134–13143.

[109] Hasan, M.K., Rahman, W., Bagher Zadeh, A., Zhong, J., Tanveer, M.I., Morency, L.P., Hoque, M.E., 2019. UR-FUNNY: A multimodal language dataset for understanding humor, in: EMNLP-IJCNLP.

[110] Haurilet, M., Roitberg, A., Stiefelhagen, R., 2019. It's not about the journey; it's about the destination: Following soft paths under question-guidance for visual reasoning, in: CVPR.

[111] Hazarika, D., Zimmermann, R., Poria, S., 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. arXiv preprint arXiv:2005.03545 .

[112] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. CVPR , 770–778.

[113] Hendricks, L.A., Hu, R., Darrell, T., Akata, Z., 2018. Generating counterfactual explanations with natural language. ArXiv abs/1806.09809.

[114] Herdade, S., Kappeler, A., Boakye, K., Soares, J., 2019. Image captioning: Transforming objects into words, in: NeurIPS, pp. 11137–11147.

[115] Hessel, J., Lee, L., 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think!

[116] 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: NIPS.

[117] Hinz, T., Heinrich, S., Wermter, S., 2019. Generating multiple objects at spatially distinct locations. ArXiv abs/1901.00686.

[118] Hirasawa, T., Komachi, M., 2019. Debiasing word embeddings improves multimodal machine translation. ArXiv abs/1905.10464.

[119] Hirasawa, T., Yamagishi, H., Matsumura, Y., Komachi, M., 2019. Multimodal machine translation with embedding prediction. ArXiv abs/1904.00639.

[120] Hodosh, M., Young, P., Hockenmaier, J., 2013. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). J. Artif. Intell. Res. 47, 853–899.

[121] Hong, S., Yang, D., Choi, J., Lee, H., 2018. Inferring semantic layout for hierarchical text-to-image synthesis. CVPR , 7986–7994.

[122] Hong, Y., Rodriguez-Opazo, C., Wu, Q., Gould, S., 2020. Sub-instruction aware vision-and-language navigation. ArXiv abs/2004.02707.

[123] Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y., 2019a. Joint syntax representation learning and visual cue translation for video captioning. ICCV , 8917–8926.

[124] Hou, M., Tang, J., Zhang, J., Kong, W., Zhao, Q., 2019b. Deep multimodal multilinear fusion with high-order polynomial pooling, in: NeurIPS, pp. 12136–12145.

[125] Hu, D., Wang, C., Nie, F., Li, X., 2019. Dense multimodal fusion for hierarchically joint representation, in: ICASSP, pp. 3941–3945.

[126] Hu, J., Shen, L., Sun, G., 2018a. Squeeze-and-excitation networks. CVPR , 7132–7141.

[127] Hu, R., Andreas, J., Darrell, T., Saenko, K., 2018b. Explainable neural computation via stack neural module networks. ArXiv abs/1807.08556.

[128] Huang, H., Jain, V., Mehta, H., Ku, A., Magalhães, G., Baldridge, J., Ie, E., 2019a. Transferable representation learning in vision-and-language navigation. ICCV , 7403–7412.

[129] Huang, J., Alfadly, M., Ghanem, B., Worring, M., 2019b. Assessing the robustness of visual question answering. ArXiv abs/1912.01452.

[130] Huang, L., Wang, W., Xia, Y., Chen, J., 2019c. Adaptively aligned image captioning via adaptive attention time, in: NeurIPS, pp. 8942–8951.

[131] Huang, P., Huang, J., Guo, Y., Qiao, M., Zhu, Y., 2019d. Multi-grained attention with object-level grounding for visual question answering, in: ACL.

[132] Huang, P.Y., Hu, J., Chang, X., Hauptmann, A.G., 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting, in: ACL.

[133] Huang, P.Y., Liu, F., Shiang, S.R., Oh, J., Dyer, C., 2016. Attention-based multimodal neural machine translation, in: First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 639–645.

[134] Hudson, D.A., Manning, C.D., 2019. Gqa: a new dataset for compositional question answering over real-world images. ArXiv abs/1902.09506.

[135] Iashin, V., Rahtu, E., 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. ArXiv abs/2005.08271.

[136] Ive, J., Madhyastha, P., Specia, L., 2019. Distilling translations with visual awareness. ArXiv abs/1906.07701.

[137] Jain, U., Lazebnik, S., Schwing, A.G., 2018. Two can play this game: Visual dialog with discriminative question generation and answering. CVPR , 5754–5763.

[138] Jain, U., Zhang, Z., Schwing, A.G., 2017. Creativity: Generating diverse questions using variational autoencoders. CVPR , 5415–5424.

[139] Ji, R., Cao, D., Lin, D., 2015. Cross-modality sentiment analysis for social multimedia, in: IEEE BigMM, pp. 28–31.

[140] Ji, R., Cao, D., Lin, D., 2015. Cross-modality sentiment analysis for social multimedia. BigMM , 28–31.

[141] Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., Diesner, J., Gao, J., 2019. Tiger: Text-to-image grounding for image caption evaluation. ArXiv abs/1909.02050.

[142] Johnson, J.E., Gupta, A., Fei-Fei, L., 2018. Image generation from scene graphs. CVPR , 1219–1228.

[143] Kafle, K., Kanan, C., 2017. An analysis of visual question answering algorithms. ICCV , 1983–1991.

[144] Kafle, K., Shrestha, R., Cohen, S., Price, B., Kanan, C., 2020. Answering questions about data visualizations using efficient bimodal fusion, in: WACV.

[145] Kafle, K., Shrestha, R., Kanan, C., 2019. Challenges and prospects in vision and language research. Frontiers in Artificial Intelligence .

[146] Kanehira, A., Takemoto, K., Inayoshi, S., Harada, T., 2019. Multimodal explanations by predicting counterfactuality in videos. CVPR , 8586–8594.

[147] Kang, G.C., Lim, J., Zhang, B.T., 2019. Dual attention networks for visual reference resolution in visual dialog. ArXiv abs/1902.09368.

[148] Karpathy, A., Fei-Fei, L., 2017. Deep visual-semantic alignments for generating image descriptions. IEEE TPAMI 39, 664–676.

[149] Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J.,

Choi, Y., Srinivasa, S.S., 2019a. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. CVPR , 6734–6742.

[150] Ke, L., Pei, W., Li, R., Shen, X., Tai, Y.W., 2019b. Reflective decoding network for image captioning. ICCV , 8887–8896.

[151] Kiela, D., Bhooshan, S., Firooz, H., Testuggine, D., 2019. Supervised multimodal bitransformers for classifying images and text. ArXiv abs/1909.02950.

[152] Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., Erdem, E., 2017. Re-evaluating automatic metrics for image captioning. ArXiv abs/1612.07600.

[153] Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S., 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning. CVPR , 6264–6273.

[154] Kim, H., Bansal, M., 2019. Improving visual question answering by referring to generated paragraph captions, in: ACL.

[155] Kim, J.H., Parikh, D., Batra, D., Zhang, B.T., Tian, Y., 2017. Codraw: Visual dialog for collaborative drawing. ArXiv abs/1712.05558.

[156] Kingma, D.P., Dhariwal, P., 2018. Glow: Generative flow with invertible 1x1 convolutions. ArXiv abs/1807.03039.

[157] Kiros, R., Salakhutdinov, R., Zemel, R., 2014. Unifying visual-semantic embeddings with multimodal neural language models. ArXiv abs/1411.2539.

[158] Klein, T., Nabi, M., 2019. Attention is (not) all you need for commonsense reasoning, in: ACL.

[159] Kornuta, T., Rajan, D., Shivade, C., Asseman, A., Ozcan, A.S., 2019. Leveraging medical visual question answering with supporting facts. ArXiv abs/1905.12008.

[160] Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S., 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. ArXiv abs/2004.02857.

[161] Krause, J., Johnson, J.E., Krishna, R., Fei-Fei, L., 2017. A hierarchical approach for generating descriptive image paragraphs. CVPR , 3337–3345.

[162] Krishna, R., Bernstein, M., Fei-Fei, L., 2019. Information maximizing visual question generation. CVPR , 2008–2018.

[163] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L., 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 123, 32–73.

[164] Kumar, A., Mittal, T., Manocha, D., 2020. Mcqa: Multimodal co-attention based network for question answering. ArXiv abs/2004.12238.

[165] Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q., 2015. From word embeddings to document distances, in: ICML, p. 957–966.

[166] Laina, I., Rupprecht, C., Navab, N., 2019. Towards unsupervised image captioning with shared multimodal embeddings. ICCV , 7413–7423.

[167] Landi, F., Baraldi, L., Cornia, M., Corsini, M., Cucchiara, R., 2019a. Perceive, transform, and act: Multi-modal attention networks for vision-and-language navigation. ArXiv abs/1911.12377.

[168] Landi, F., Baraldi, L., Corsini, M., Cucchiara, R., 2019b. Embodied vision-and-language navigation with dynamic convolutional filters, in: BMVC.

[169] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

[170] Lee, S.W., Gao, T., Yang, S., Yoo, J., Ha, J.W., 2019. Large-scale answerer in questioner's mind for visual dialog question generation. ArXiv abs/1902.08355.

[171] Lei, J., Wang, L., Shen, Y., Yu, D., Berg, T.L., Bansal, M., 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. ArXiv abs/2005.05402.

[172] Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M., 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, in: AAAI.

[173] Li, J., Wang, X., Tang, S., Shi, H., Wu, F., Zhuang, Y., Wang, W.Y., 2020b. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. CVPR , 12120–12129.

[174] Li, L., Gan, Z., Cheng, Y., Liu, J., 2019a. Relation-aware graph attention network for visual question answering, in: ICCV.

[175] Li, L., Gong, B., 2019. End-to-end video captioning with multitask reinforcement learning. WACV , 339–348.

[176] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W., 2019b. Visualbert: A simple and performant baseline for vision and language. ArXiv abs/1908.03557.

[177] Li, S., Xiao, T., Li, H., Yang, W., Wang, X., 2017. Identity-aware textual-visual matching with latent co-attention. ICCV , 1908–1917.

[178] Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J., 2019c. Object-driven text-to-image synthesis via adversarial training. CVPR , 12166–12174.

[179] Li, X., Li, C., Xia, Q., Bisk, Y., Çelikyilmaz, A., Gao, J., Smith, N.A., Choi, Y., 2019d. Robust navigation with language pretraining and stochastic sampling, in: EMNLP/IJCNLP.

[180] Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C., 2019e. Beyond rnns: Positional self-attention with co-attention for video question answering, in: AAAI.

[181] Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J., 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. ArXiv abs/2004.06165.

[182] Li, Y., Duan, N., Zhou, B., Chu, X.R., Ouyang, W., Wang, X., 2018a. Visual question generation as dual task of visual question answering. CVPR , 6116–6124.

[183] Li, Y., Min, M.R., Shen, D., Carlson, D.E., Carin, L., 2018b. Video generation from text, in: AAAI.

[184] Li, Y., Yang, Y., Wang, J., Xu, W., 2018c. Zero-shot transfer vqa dataset. ArXiv abs/1811.00692.

[185] Li, Z., Li, Y., Lu, H., 2019f. Improve image captioning by self-attention, in: NeurIPS, pp. 91–98.

[186] Li, Z., Tran, Q.H., Mai, L., Lin, Z., Yuille, A.L., 2020d. Context-aware group captioning via self-attention and contrastive features. ArXiv abs/2004.03708.

[187] Liang, J., Jiang, L., Cao, L., Kalantidis, Y., Li, L.J., Hauptmann, A.G., 2019a. Focal visual-text attention for memex question answering. IEEE TPAMI 41, 1893–1908.

[188] Liang, P.P., Liu, Z., Tsai, Y.H., Zhao, Q., Salakhutdinov, R., Morency, L., 2019b. Learning representations from imperfect time series data via tensor rank regularization, in: ACL, pp. 1569–1576.

[189] Libovický, J., Helcl, J., 2017. Attention strategies for multi-source sequence-to-sequence learning, in: ACL.

[190] Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries, in: ACL 2004.

[191] Lin, J., Jain, U., Schwing, A.G., 2019. Tab-vcr: Tags and attributes based visual commonsense reasoning baselines. CVPR .

[192] Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. ArXiv abs/1405.0312.

[193] Liu, L., Tang, J., Wan, X., Guo, Z., 2019a. Generating diverse and descriptive image captions using visual paraphrases. ICCV , 4239–4248.

[194] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019b. Roberta: A robustly optimized bert pretraining approach. ArXiv abs/1907.11692.

[195] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Bagher Zadeh, A., Morency, L.P., 2018. Efficient low-rank multimodal fusion with modality-specific factors, in: ACL.

[196] Lu, J., Batra, D., Parikh, D., Lee, S., 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. ArXiv abs/1908.02265.

[197] Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S., 2019b. 12-in-1: Multi-task vision and language representation learning. ArXiv abs/1912.02315.

[198] Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering. ArXiv abs/1606.00061.

[199] Luong, T., Pham, H., Manning, C.D., 2015. Effective approaches to

attention-based neural machine translation, in: EMNLP.

[200] Ma, C.Y., Lu, J., Wu, Z., Al-Regib, G., Kira, Z., Socher, R., Xiong, C., 2019a. Self-monitoring navigation agent via auxiliary progress estimation. ArXiv abs/1901.03035.

[201] Ma, C.Y., Wu, Z., Al-Regib, G., Xiong, C., Kira, Z., 2019b. The regretful agent: Heuristic-aided navigation through progress estimation. CVPR , 6725–6733.

[202] Mai, S., Hu, H., Xing, S., 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing, in: ACL, pp. 481–492.

[203] Mai, S., Hu, H., Xing, S., 2020a. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in: AAAI.

[204] Mai, S., Xing, S., Hu, H., 2020b. Locally confined modality fusion network with a global perspective for multimodal human affective computing. Trans. Multi. 22, 122–137.

[205] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S., 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. ArXiv abs/1806.06228.

[206] Manjunatha, V., Saini, N., Davis, L., 2019. Explicit bias discovery in visual question answering models. CVPR , 9554–9563.

[207] Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H., 2019. Mode seeking generative adversarial networks for diverse image synthesis. CVPR , 1429–1437.

[208] Massiceti, D., Siddharth, N., Dokania, P.K., Torr, P.H.S., 2018. Flipdial: A generative model for two-way visual dialogue. CVPR , 6097–6105.

[209] Mathews, A.P., Xie, L., He, X., 2016. Senticap: Generating image descriptions with sentiments. ArXiv abs/1510.01431.

[210] Mathews, A.P., Xie, L., He, X., 2018. Semstyle: Learning to generate stylised image captions using unaligned text. CVPR , 8591–8600.

[211] Mazaheri, A., Shah, M., 2018. Visual text correction, in: ECCV.

[212] Mei, T., Zhang, W., Yao, T., 2020. Vision and language: from visual perception to content creation. APSIPA Transactions on Signal and Information Processing .

[213] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. ArXiv abs/1310.4546.

[214] Mirowski, P.W., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., Hadsell, R., 2017. Learning to navigate in complex environments. ArXiv abs/1611.03673.

[215] Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A., 2019. Ocr-vqa: Visual question answering by reading text in images, in: ICDAR, pp. 947–952.

[216] Mittal, G., Marwah, T., Balasubramanian, V.N., 2017. Sync-draw: Automatic video generation using deep recurrent attentive architectures. ACM MM .

[217] Miyazawa, K., Aoki, T., Horii, T., Nagai, T., 2020. lambert: Language and action learning using multimodal bert. ArXiv abs/2004.07093.

[218] Mogadala, A., Kalimuthu, M., Klakow, D., 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. ArXiv .

[219] Monkaresi, H., Hussain, M.S., Calvo, R.A., 2012. Classification of affects using head movement, skin color features and physiological signals, in: SMC, pp. 2664–2669.

[220] Mudrakarta, P.K., Taly, A., Sundararajan, M., Dhamdhere, K., 2018. Did the model understand the question? arXiv preprint arXiv:1805.05492 .

[221] Murahari, V.S., Batra, D., Parikh, D., Das, A., 2019. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. ArXiv abs/1912.02379.

[222] Nagarajan, T., Grauman, K., 2018. Attributes as operators. ArXiv abs/1803.09851.

[223] Nguyen, D.K., Okatani, T., 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. CVPR , 6087–6096.

[224] Nguyen, D.K., Okatani, T., 2019. Multi-task learning of hierarchical vision-language representation. CVPR , 10484–10493.

[225] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R., 2020. Counterfactual vqa: A cause-effect look at language bias. ArXiv abs/2006.04315.

[226] Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., Wen, J.R., 2019. Recursive visual attention in visual dialog, in: CVPR.

[227] Norcliffe-Brown, W., Vafeias, E., Parisot, S., 2018. Learning conditioned graph structures for interpretable visual question answering, in: NeurIPS.

[228] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention u-net: Learning where to look for the pancreas. ArXiv abs/1804.03999.

[229] Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C., 2020a. Spatio-temporal graph for video captioning with knowledge distillation. ArXiv abs/2003.13942.

[230] Pan, J., Goyal, Y., Lee, S., 2019. Question-conditioned counterfactual image generation for vqa. ArXiv abs/1911.06352.

[231] Pan, Y., Qiu, Z., Yao, T., Li, H., Mei, T., 2017. To create what you tell: Generating videos from captions. ACM MM .

[232] Pan, Y., Yao, T., Li, Y., Mei, T., 2020b. X-linear attention networks for image captioning. ArXiv abs/2003.14080.

[233] Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation, in: ACL.

[234] Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M., 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. CVPR , 8779–8788.

[235] Patro, B., Patel, S., Namboodiri, V., 2020a. Robust explanations for visual question answering, in: WACV.

[236] Patro, B.N., Anupriy, Namboodiri, V.P., 2019. Probabilistic framework for solving visual dialog. ArXiv abs/1909.04800.

[237] Patro, B.N., Kumar, S., Kurmi, V.K., Namboodiri, V., 2018. Multimodal differential network for visual question generation, in: EMNLP.

[238] Patro, B.N., Pate, S., Namboodiri, V.P., 2020b. Robust explanations for visual question answering. WACV , 1566–1575.

[239] Peña, A., Serna, I., Morales, A., Fiérrez, J., 2020. Bias in multimodal AI: testbed for fair automatic recruitment, in: CVPR Workshops, pp. 129–137.

[240] Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global vectors for word representation, in: EMNLP, pp. 1532–1543.

[241] Pérez-Rosas, V., Mihalcea, R., Morency, L.P., 2013. Utterance-level multimodal sentiment analysis, in: ACL.

[242] Pérez-Rúa, J.M., Vielzeuf, V., Pateux, S., Baccouche, M., Jurie, F., 2019. Mfas: Multimodal fusion architecture search. CVPR , 6959–6968.

[243] Pham, H., Liang, P.P., Manzini, T., Morency, L., Póczos, B., 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities, in: AAAI, pp. 6892–6899.

[244] Plummer, B.A., Wang, L., Cervantes, C., Caicedo, J.C., Hockenmaier, J., Lazebnik, S., 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV 123, 74–93.

[245] Pollard, A.E., Shapiro, J.L., 2020. Visual question answering as a multi-task problem. ArXiv abs/2007.01780.

[246] Poria, S., Cambria, E., Gelbukh, A., 2015a. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: EMNLP.

[247] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P., 2017. Context-dependent sentiment analysis in user-generated videos, in: ACL.

[248] Poria, S., Cambria, E., Hussain, A., Huang, G.B., 2015b. Towards an intelligent framework for multimodal affective data analysis. Neural Networks 63, 104–16.

[249] Poria, S., Chaturvedi, I., Cambria, E., Hussain, A., 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: in ICDM, pp. 439–448.

[250] Poria, S., Chaturvedi, I., Cambria, E., Hussain, A., 2016. Convo-

lutional mkl based multimodal emotion recognition and sentiment analysis. ICDM , 439–448.

[251] Poria, S., Hussain, A., Cambria, E., 2018. Multimodal sentiment analysis, in: Socio-Affective Computing.

[252] Qi, D., Su, L., Song, J., Cui, E.D.B., Bharti, T., Sacheti, A., 2020a. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. ArXiv abs/2001.07966.

[253] Qi, F., Yang, X., Xu, C., 2018. A unified framework for multimodal domain adaptation, in: ACM MM, p. 429–437.

[254] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d., 2020b. Reverie: Remote embodied visual referring expression in real indoor environments, in: CVPR.

[255] Qiao, T., Zhang, J., Xu, D., Tao, D., 2019. Mirrorgan: Learning text-to-image generation by redescription. CVPR , 1505–1514.

[256] Qin, Y., Du, J., Zhang, Y., Lu, H., 2019. Look back and predict forward in image captioning. CVPR , 8359–8367.

[257] Rahman, T., Xu, B., Sigal, L., 2019. Watch, listen and tell: Multimodal weakly supervised dense event captioning. ICCV , 8907–8916.

[258] Rajagopalan, S.S., Morency, L.P., Baltrusaitis, T., Goecke, R., 2016. Extending long short-term memory for multi-view structured learning, pp. 338–353.

[259] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J., 2019. Stand-alone self-attention in vision models. ArXiv abs/1906.05909.

[260] Ramakrishnan, S., Agrawal, A., Lee, S., 2018. Overcoming language priors in visual question answering with adversarial regularization. ArXiv abs/1810.03649.

[261] Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A., 2016. You only look once: Unified, real-time object detection. CVPR , 779–788.

[262] Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H., 2016a. Learning what and where to draw, in: NIPS.

[263] Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016b. Generative adversarial text to image synthesis. ArXiv abs/1605.05396.

[264] Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE TPAMI 39, 1137–1149.

[265] Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. CVPR , 1179–1195.

[266] Sahu, G., Vechtomova, O., 2019. Dynamic fusion for multimodal data. ArXiv abs/1911.03821.

[267] Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. ArXiv abs/1606.03498.

[268] Samek, W., Müller, K.R., 2019. Towards explainable artificial intelligence, in: Explainable AI.

[269] Sammani, F., Melas-Kyriazi, L., 2020. Show, edit and tell: A framework for editing image captions. ArXiv abs/2003.03107.

[270] Sarkar, C., Bhatia, S., Agarwal, A., Li, J., 2014. Feature analysis for computational personality recognition using youtube personality data set, in: WCPR, p. 11–14.

[271] Schwartz, I., Yu, S., Hazan, T., Schwing, A.G., 2019. Factor graph attention. CVPR , 2039–2048.

[272] Seo, P.H., Lehrmann, A., Han, B., Sigal, L., 2017. Visual reference resolution using attention memory for visual dialog, in: Advances in Neural Information Processing Systems 30, pp. 3719–3729.

[273] Seo, P.H., Sharma, P., Levinboim, T., Han, B., Soricut, R., 2020. Reinforcing an image caption generator using off-line human feedback, in: AAAI.

[274] Sharif, N., White, L., Bennamoun, M., Shah, S.A.A., 2018. Nneval: Neural network based evaluation metric for image captioning, in: ECCV.

[275] Sharma, S., Suhubdy, D., Michalski, V., Kahou, S.E., Bengio, Y., 2018a. Chatpainter: Improving text to image generation using dialogue. ArXiv abs/1802.08216.

[276] Sharma, V., Kalra, A., Vaibhav, Chaudhary, S., Patel, L., Morency, L.P., 2018b. Attend and attack : Attention guided adversarial attacks on visual question answering models.

[277] Shen, T., Kar, A., Fidler, S., 2019. Learning to caption images through a lifetime by asking questions. ICCV , 10392–10401.

[278] Shi, B., Ji, L., Liang, Y., Duan, N., Chen, P., Niu, Z., Zhou, M., 2019a. Dense procedure captioning in narrated instructional videos, in: ACL.

[279] Shi, J., Zhang, H., Li, J.Z., 2019b. Explainable and explicit visual reasoning over scene graphs. CVPR , 8368–8376.

[280] Shi, L., Shuang, K., Geng, S., Su, P., Jiang, Z., Gao, P., Fu, Z., de Melo, G., Su, S., 2020a. Contrastive visual-linguistic pretraining. ArXiv abs/2007.13135.

[281] Shi, Z., Zhou, X., Qiu, X., Zhu, X., 2020b. Improving image captioning with better use of captions. arXiv:2006.11807.

[282] Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J., 2019. Engaging image captioning via personality. CVPR , 12508–12518.

[283] Siddiquie, B., Chisholm, D., Divakaran, A., 2015. Exploiting multimodal affect and semantics to identify politically persuasive web videos, in: ICMI.

[284] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

[285] Sindagi, V., Patel, V.M., 2020. Ha-ccn: Hierarchical attention-based crowd counting network. IEEE TIP 29, 323–335.

[286] Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. ArXiv abs/2005.13799.

[287] Sinha, A., Dolz, J.E., 2019. Multi-scale guided attention for medical image segmentation. ArXiv abs/1906.02849.

[288] Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., Shen, H.T., 2017. Hierarchical lstm with adjusted temporal attention for video captioning. ArXiv abs/1706.01231.

[289] Song, M., Bu, J., Chen, C., Li, N., 2004. Audio-visual based emotion recognition - a new approach. CVPR 2, II–II.

[290] Srivastava, N., Salakhutdinov, R.R., 2012. Multimodal learning with deep boltzmann machines, in: NIPS, pp. 2222–2230.

[291] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2020. Vlbert: Pre-training of generic visual-linguistic representations. ArXiv abs/1908.08530.

[292] Su, Y., Fan, K., Bach, N., Kuo, C.C.J., Huang, F., 2019. Unsupervised multi-modal neural machine translation. CVPR , 10474–10483.

[293] Suhr, A., Lewis, M., Yeh, J., Artzi, Y., 2017. A corpus of natural language for visual reasoning, in: ACL.

[294] Suhr, A., Zhou, S., Zhang, I.D., Bai, H., Artzi, Y., 2019. A corpus for reasoning about natural language grounded in photographs. ArXiv abs/1811.00491.

[295] Sun, C., Baradel, F., Murphy, K., Schmid, C., 2019a. Contrastive bidirectional transformer for temporal representation learning. ArXiv abs/1906.05743.

[296] Sun, C., Gan, C., Nevatia, R., 2015. Automatic concept discovery from parallel text and visual corpora. ICCV , 2596–2604.

[297] Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C., 2019b. Videobert: A joint model for video and language representation learning. ICCV , 7463–7472.

[298] Sun, Z., Sarma, P.K., Sethares, W.A., Liang, Y., 2019c. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. CoRR abs/1911.05544.

[299] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. ArXiv abs/1409.3215.

[300] Tan, F., Feng, S., Ordonez, V., 2018. Text2scene: Generating abstract scenes from textual descriptions. ArXiv abs/1809.01110.

[301] Tan, F., Feng, S., Ordonez, V., 2019a. Text2scene: Generating compositional scenes from textual descriptions. CVPR , 6703–6712.

[302] Tan, H., Bansal, M., 2018. Object ordering with bidirectional matchings for visual reasoning. ArXiv abs/1804.06870.

[303] Tan, H., Yu, L., Bansal, M., 2019b. Learning to navigate unseen environments: Back translation with environmental dropout. ArXiv abs/1904.04195.

[304] Tan, H.H., Bansal, M., 2019. Lxmert: Learning cross-modality encoder representations from transformers. ArXiv abs/1908.07490.

[305] Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W., 2019. Learning to compose dynamic tree structures for visual contexts. CVPR , 6612–6621.

[306] Tang, R., Ma, C., Zhang, W., Wu, Q., Yang, X., 2020. Semantic equivalent adversarial data augmentation for visual question answering. ArXiv abs/2007.09592.

[307] Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S., 2016. Movieqa: Understanding stories in movies through question-answering. CVPR , 4631–4640.

[308] Tay, Y., Luu, A.T., Hui, S.C., 2018. Hermitian co-attention networks for text matching in asymmetrical domains, in: in IJCAI.

[309] Teney, D., Hengel, A.V.D., 2016. Zero-shot visual question answering. ArXiv abs/1611.05546.

[310] Thomee, B., Lew, M.S., 2012. Interactive search in image retrieval: a survey. IJMIR 1, 71–86.

[311] Tian, Y., Guan, C., Goodman, J., Moore, M., Xu, C., 2019. Audio-visual interpretable and controllable video captioning, in: CVPR Workshops.

[312] Tommasi, T., Mallya, A., Plummer, B.A., Lazebnik, S., Berg, A.C., Berg, T.L., 2016. Solving visual madlibs with multiple cues. ArXiv abs/1608.03410.

[313] Tran, A., Mathews, A.P., Xie, L., 2020. Transform and tell: Entity-aware news image captioning. ArXiv abs/2004.08070.

[314] Trinh, T.H., Luong, M.T., Le, Q.V., 2019. Selfie: Self-supervised pretraining for image embedding. ArXiv abs/1906.02940.

[315] Trott, A., Xiong, C., Socher, R., 2018. Interpretable counting for visual question answering. ArXiv abs/1712.08697.

[316] Tsai, Y.H., Liang, P.P., Zadeh, A., Morency, L., Salakhutdinov, R., 2019a. Learning factorized multimodal representations, in: in ICLR.

[317] Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019b. Multimodal transformer for unaligned multimodal language sequences, in: ACL, pp. 6558–6569.

[318] Tschannen, M., Bachem, O., Lucic, M., 2018. Recent advances in autoencoder-based representation learning. ArXiv abs/1812.05069.

[319] Udandarao, V., Maiti, A., Srivatsav, D., Vyalla, S.R., Yin, Y., Shah, R.R., 2020. Cobra: Contrastive bi-modal representation algorithm. ArXiv abs/2005.03687.

[320] Uppal, S., Madan, A., Bhagat, S., Yu, Y., Shah, R.R., 2020. C3vqg: Category consistent cyclic visual question generation. ArXiv abs/2005.07771.

[321] Vahdat, A., Kautz, J., 2020. Nvae: A deep hierarchical variational autoencoder. ArXiv abs/2007.03898.

[322] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. ArXiv abs/1706.03762.

[323] Vedantam, R., Desai, K., Lee, S., Rohrbach, M., Batra, D., Parikh, D., 2019. Probabilistic neural-symbolic models for interpretable visual question answering, in: ICML.

[324] Vedantam, R., Zitnick, C.L., Parikh, D., 2015. Cider: Consensus-based image description evaluation. CVPR , 4566–4575.

[325] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., Saenko, K., 2015. Sequence to sequence – video to text. ICCV , 4534–4542.

[326] Vered, G., Oren, G., Atzmon, Y., Chechik, G., 2019. Joint optimization for cooperative image captioning. ICCV , 8897–8906.

[327] Verma, G., Vinay, V., Bansal, S., Oberoi, S., Sharma, M., Gupta, P., 2020. Using image captions and multitask learning for recommending query reformulations. Advances in Information Retrieval 12035, 681 – 696.

[328] Vijaikumar, M., Shevade, S., Narasimha Murty, M., 2020. Gamma: A graph and multi-view memory attention mechanism for top-n heterogeneous recommendation, in: Advances in Knowledge Discovery and Data Mining, pp. 28–40.

[329] Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. CVPR .

[330] Vo, N.S., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J., 2019. Composing text and image for image retrieval - an empirical odyssey. CVPR , 6432–6441.

[331] Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W., 2019a. Controllable video captioning with pos sequence guidance based on gated fusion network. ICCV , 2641–2650.

[332] Wang, C., Yang, H., Meinel, C., 2018a. Image captioning with deep bidirectional lstms and multi-task learning. ACM MM 14.

[333] Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y., 2018b. Bidirectional attentive fusion with context gating for dense video captioning. CVPR , 7190–7198.

[334] Wang, J., Wang, W., Huang, Y., Wang, L., Tan, T., 2018c. M3: Multimodal memory modelling for video captioning. CVPR , 7512–7520.

[335] Wang, L., Bai, Z., Zhang, Y., Lu, H., 2020a. Show, recall, and tell: Image captioning with recall mechanism. ArXiv abs/2001.05876.

[336] Wang, Q., Chan, A.B., 2018. Gated hierarchical attention for image captioning, in: ICCV.

[337] Wang, W., Tran, D., Feiszli, M., 2019b. What makes training multi-modal networks hard? ArXiv abs/1905.12681.

[338] Wang, X., Huang, Q., Çelikyilmaz, A., Gao, J., Shen, D., fang Wang, Y., Wang, W.Y., Zhang, L., 2019c. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. CVPR , 6622–6631.

[339] Wang, X., Jain, V., Ie, E., Wang, W.Y., Kozareva, Z., Ravi, S., 2020b. Environment-agnostic multitask learning for natural language grounded navigation. ArXiv abs/2003.00443.

[340] Wang, X., Liu, Y., Shen, C., Ng, C.C., Luo, C., Jin, L., Chan, C.S., van den Hengel, A., Wang, L., 2020c. On the general value of evidence, and bilingual scene-text visual question answering. ArXiv abs/2002.10215.

[341] Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., Wang, W.Y., 2019d. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. ICCV , 4580–4590.

[342] Wang, X., Xiong, W., Wang, H., Wang, W.Y., 2018d. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. ArXiv abs/1803.07729.

[343] Wang, Y., Joty, S.R., Lyu, M.R., King, I., Xiong, C., Hoi, S.C.H., 2020d. Vd-bert: A unified vision and dialog transformer with bert. ArXiv abs/2004.13278.

[344] Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L., 2019e. Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: AAAI, pp. 7216–7223.

[345] Wang, Y., Wang, S., Tang, J., O'Hare, N., Chang, Y., Li, B., 2016. Hierarchical attention network for action recognition in videos. ArXiv abs/1607.06416.

[346] Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J., 2019f. Camp: Cross-modal adaptive message passing for text-image retrieval. ICCV , 5763–5772.

[347] Wu, A., Zhu, L., Han, Y., Yang, Y., 2019a. Connective cognition network for directional visual commonsense reasoning, in: NeurIPS, pp. 5669–5679.

[348] Wu, J., Hu, Z., Mooney, R.J., 2019b. Generating question relevant captions to aid visual question answering, in: ACL.

[349] Wu, J., Mooney, R., 2019a. Self-critical reasoning for robust visual question answering, in: NeurIPS, pp. 8604–8614.

[350] Wu, J., Mooney, R.J., 2019b. Self-critical reasoning for robust visual question answering. ArXiv abs/1905.09998.

[351] Wu, L., Ge, Y., Liu, Q., Chen, E., Hong, R., Wang, M., Du, J., 2018. Explainable social contextual image recommendation with hierarchical attention. ArXiv abs/1806.00723.

[352] Wu, Q., Wang, P., Shen, C., Dick, A.R., van den Hengel, A., 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. CVPR , 4622–4630.

[353] Wu, Y., Wang, S., Song, G., Huang, Q., 2019c. Learning fragment self-attention embeddings for image-text matching, in: ACM MM, p. 2088–2096.

[354] Xia, Q., Li, X., Li, C., Bisk, Y., Sui, Z., Choi, Y., Smith, N.A., 2020. Multi-view learning for vision-and-language navigation. ArXiv abs/2003.00857.

[355] Xie, N., Ras, G., Gerven, M.V., Doran, D., 2020. Explainable deep learning: A field guide for the uninitiated. ArXiv .

[356] Xiong, C., Merity, S., Socher, R., 2016. Dynamic memory networks for visual and textual question answering, in: ICML.

[357] Xiong, C., Zhong, V., Socher, R., 2017. Dynamic coattention networks for question answering. ArXiv abs/1611.01604.

[358] Xu, H., Saenko, K., 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. ArXiv abs/1511.05234.

[359] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. ArXiv abs/1502.03044.

[360] 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. CVPR , 1316–1324.

[361] Xu, X., Chen, X., Liu, C., Rohrbach, A., Darrell, T., Song, D., 2017. Can you fool ai with adversarial examples on a visual turing test? ArXiv abs/1709.08693.

[362] Xu, Y., Chen, L., Cheng, Z., Duan, L., Luo, J., 2019a. Open-ended visual question answering by multi-modal domain adaptation. ArXiv abs/1911.04058.

[363] Xu, Y., Wu, B., Shen, F., Fan, Y., Zhang, Y., Shen, H.T., Liu, W., 2019b. Exact adversarial attack to image captioning via structured output learning with latent variables. CVPR , 4130–4139.

[364] Yang, M., Zhao, W., Xu, W., Feng, Y., Zhao, Z., Chen, X., Lei, K., 2019. Multitask learning for cross-domain image captioning. IEEE Transactions on Multimedia 21, 1047–1061.

[365] Yang, P., Chen, B., Zhang, P., Sun, X., 2020a. Visual agreement regularized training for multi-modal machine translation, in: AAAI.

[366] Yang, X., Tang, K., Zhang, H., Cai, J., 2019a. Auto-encoding scene graphs for image captioning. CVPR , 10677–10686.

[367] Yang, X., Zhang, H., Cai, J., 2019b. Learning to collocate neural modules for image captioning. ICCV , 4249–4259.

[368] Yang, Y., Li, Y., Fermüller, C., Aloimonos, Y., 2015. Neural self talk: Image understanding via continuous questioning and answering. ArXiv abs/1512.03460.

[369] Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., Takemura, H., 2020b. Bert representations for video question answering. WACV , 1545–1554.

[370] Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., Takemura, H., 2020c. Bert representations for video question answering, in: WACV.

[371] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J., 2016a. Stacked attention networks for image question answering. CVPR , 21–29.

[372] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., Hovy, E.H., 2016b. Hierarchical attention networks for document classification, in: HLT-NAACL.

[373] Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., Cohen, W.W., 2016c. Encode, review, and decode: Reviewer module for caption generation. ArXiv abs/1605.07912.

[374] Yao, T., Pan, Y., Li, Y., Mei, T., 2017. Incorporating copying mechanism in image captioning for learning novel objects. CVPR , 5263–5271.

[375] Yao, T., Pan, Y., Li, Y., Mei, T., 2019. Hierarchy parsing for image captioning. ICCV , 2621–2629.

[376] Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J., 2019a. Semantics disentangling for text-to-image generation. CVPR , 2322–2331.

[377] Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., 2019b. Context and attribute grounded dense captioning. CVPR , 6234–6243.

[378] Yin, X., Ordonez, V., Feng, S., 2019c. Chat-crowd: A dialog-based platform for visual layout composition. ArXiv abs/1812.04081.

[379] Yu, F., Deng, Z., Narasimhan, K., Russakovsky, O., 2020a. Take the scenic route: Improving generalization in vision-and-language navigation. CVPR Workshops , 4000–4004.

[380] Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H., 2020b. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. ArXiv abs/2006.16934.

[381] Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W., 2016. Video paragraph captioning using hierarchical recurrent neural networks. CVPR , 4584–4593.

[382] Yu, W., Zhou, J., Yu, W., Liang, X., Xiao, N., 2019a. Heterogeneous graph learning for visual commonsense reasoning, in: NeurIPS.

[383] Yu, Z., Cui, Y., Yu, J., Tao, D., Tian, Q., 2019b. Multimodal unified attention networks for vision-and-language interactions. ArXiv abs/1908.04107.

[384] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q., 2019c. Deep modular co-attention networks for visual question answering. CVPR , 6274–6283.

[385] Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. ICCV , 1839–1848.

[386] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P., 2017. Tensor fusion network for multimodal sentiment analysis. ArXiv abs/1707.07250.

[387] Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P., 2018a. Memory fusion network for multi-view sequential learning. ArXiv abs/1802.00927.

[388] Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P., 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: ACL.

[389] Zellers, R., Bisk, Y., Farhadi, A., Choi, Y., 2019. From recognition to cognition: Visual commonsense reasoning, in: CVPR.

[390] 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.

[391] Zeng, Z., Hu, Y., Liu, M., Fu, Y., Huang, T.S., 2006. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition, in: in MM.

[392] Zha, Z.J., Liu, J., Yang, T., Zhang, Y., 2019. Spatiotemporal-textual co-attention network for video question answering. ACM MM 15.

[393] Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C., 2019a. Raven: A dataset for relational and analogical visual reasoning. CVPR , 5312–5322.

[394] Zhang, H., Litman, D., 2018. Co-attention based neural network for source-dependent essay scoring, in: Workshop on Innovative Use of NLP for Building Educational Applications.

[395] Zhang, H., Xu, T., Li, H., 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. ICCV , 5908–5916.

[396] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2019b. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI 41, 1947–1962.

[397] Zhang, J., Wu, Q., Shen, C., Zhang, J., Lu, J., van den Hengel, A., 2018a. Asking the difficult questions: Goal-oriented visual question generation via intermediate rewards, in: ECCV.

[398] Zhang, J., Zhao, T., Yu, Z., 2018b. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog, in: SIGDIAL Conference.

[399] Zhang, S., Wang, Z., Xu, X., Guan, X., Yang, Y., 2020. Fooled by imagination: Adversarial attack to image captioning via perturbation in complex domain, in: ICME, pp. 1–6.

[400] Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J., 2020. Object relational graph with teacher-recommended learning for video captioning. ArXiv abs/2002.11566.

[401] Zhang, Z., Xie, Y., Yang, L., 2018c. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. CVPR , 6199–6208.

[402] Zhao, B., Meng, L., Yin, W., Sigal, L., 2019a. Image generation from layout. CVPR , 8576–8585.

[403] Zhao, S., Sharma, P., Levinboim, T., Soricut, R., 2019b. Informative image captioning with external sources of information. ArXiv abs/1906.08876.

[404] 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, p. 1205–1211.

[405] Zhao, W., Wu, X., Zhang, X., 2020. Memcap: Memorizing style knowledge for image captioning, in: AAAI.

[406] Zheng, C., Guo, Q., Kordjamshidi, P., 2020a. Cross-modality relevance for reasoning on language and vision. ArXiv abs/2005.06035.

[407] Zheng, W., Yan, L., Gou, C., Wang, F.Y., 2020b. Webly supervised knowledge embedding model for visual reasoning, in: CVPR.

[408] Zheng, Y., Li, Y., Wang, S., 2019. Intention oriented image captions with guiding objects. CVPR , 8387–8396.

[409] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J., 2020a. Unified vision-language pre-training for image captioning and vqa, in: AAAI.

[410] Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C., 2018a. End-to-end dense video captioning with masked transformer. CVPR , 8739–8748.

[411] Zhou, M., Cheng, R., Lee, Y.J., Yu, Z., 2018b. A visual attention grounding neural model for multimodal machine translation, in: EMNLP, pp. 3643–3653.

[412] Zhou, Y., Wang, M., Liu, D., Hu, Z., Zhang, H., 2020b. More grounded image captioning by distilling image-text matching model. ArXiv abs/2004.00390.

[413] Zhu, F., Zhu, Y., Chang, X., Liang, X., 2020a. Vision-language navigation with self-supervised auxiliary reasoning tasks. CVPR , 10009–10019.

[414] Zhu, L., Xu, Z., Yang, Y., Hauptmann, A.G., 2015. Uncovering temporal context for video question and answering. ArXiv abs/1511.04670.

[415] Zhu, M., Ahuja, A., Wei, W., Reddy, C.K., 2019a. A hierarchical attention retrieval model for healthcare question answering. WWW .

[416] Zhu, M., Pan, P., Chen, W., Yang, Y., 2019b. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. CVPR , 5795–5803.

[417] Zhu, P.F., Zhao, H., Li, X., 2020b. Dual multi-head co-attention for multi-choice reading comprehension. ArXiv abs/2001.09415.

[418] Zhu, Y., Zhu, F., Zhan, Z., Lin, B., Jiao, J., Chang, X., Liang, X., 2020c. Vision-dialog navigation by exploring cross-modal memory. CVPR , 10727–10736.