AUTHORS	ARTICLE TITLE	IDEA	YEAR
Huayu Li, Martin Renqiang Min, Yong Ge, Asim Kadav	**A Context-Aware Attention Network for Interactive Question Answering** **ABSTRACT** a new model for Interactive Question Answering (IQA), using GRUs as encoders for statements and questions, another GRU as a decoder for outputs. context-dependent word-level attention for more accurate statement representations question-guided sentence-level attention for better context modeling. accurately understands when output an answer or requires generating a supplementary question for additional input. When available, user's feedback is encoded and directly applied to update sentence-level attention to infer the answer. experiments on QA and IQA datasets demonstrate quantitatively the effectiveness with significant improvement over conventional QA models.	IQA	2017
Alane Suhr, Mike Lewis, James Yeh, Yoav Artzi	A Corpus of Natural Language for Visual Reasoning		2017
Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, Christopher Pa	A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering		
Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, Dan Jurafsky	Adversarial Learning for Neural Dialogue Generation		
Ilija Ilievski, Shuicheng Yan, Jiashi Feng	A Focused Dynamic Attention Model for Visual Question Answering		
Tanmay Gupta 1 Kevin Shih 1 Saurabh Singh 2 Derek Hoiem 1	Aligned Image-Word Representations Improve Inductive Transfer Across Vision-Language Tasks		
Aishwarya Agrawal * , Dhruv Batra †,* , Devi Parikh	Analyzing the Behavior of Visual Question Answering Models		2016
Kushal Kafle Christopher Kanan	**An Analysis of Visual Question Answering Algorithms** **Abstract** answer text-based questions about images. datasets for VQA all have flaws in both their content and the way algorithms evaluated on them. evaluation scores are inflated and determined by answering easier questions, making it difficult to compare different methods. analyze existing VQA algorithms using a new dataset. 1.6	review	2017

	million questions organized into 12 different categories. introduce questions meaningless for a given image to force a VQA system to reason about image content. propose new evaluation schemes compensate for over-represented question-types and make it easier to study the strengths and weaknesses of algorithms. analyze the performance of both baseline and state-of-the-art VQA models, including (1) multi-modal compact bilinear pooling (MCB) (2) neural module networks (3) recurrent answering units establish how attention helps certain categories more than others, determine which models work better than others, explain how simple models (e.g. MLP) can surpass more complex models (MCB) by simply learning to answer large, easy question categories.		
Santhosh K. Ramakrishnan, Ambar Pal, Gaurav Sharma, Anurag Mittal	**An Empirical Evaluation of Visual Question Answering for Novel Objects** Abstract We study the problem of answering questions about images in the harder setting, where the test questions and corresponding images contain novel objects, which were not queried about in the training data. Such setting is inevitable in real world—owing to the heavy tailed distribution of the visual categories, there would be some objects which would not be annotated in the train set. We show that the performance of two popular existing methods drop significantly(up to 28%) when evaluated on novel objects cf. known objects. We propose methods which use large existing external corpora of (i) unlabeled text, i.e. books, and (ii) images tagged with classes, to achieve novel object based visual question answering. We do systematic empirical studies, for both an oracle case where the novel objects are known textually, as well as a fully automatic case without any explicit knowledge of the novel objects, but with the minimal assumption that the novel objects are semantically related to the existing objects in training. The proposed methods for novel object based visual question answering are modular and can potentially be used with many visual question answering architectures. We show consistent improvements with the two popular architectures and give qualitative analysis of the cases where the model does well and of those where it fails to bring		

	improvements.		
Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Timothy Lillicrap	**A simple neural network module for relational reasoning** **Abstract:** Relational reasoning: central component of generally intelligent behavior, difficult for neural networks to learn. use Relation Networks (RNs) as a simple plug-and-play module to solve problems hinge on relational reasoning. tested RN-augmented networks on three tasks: (1) visual question answering using CLEVR, achieve state-of-the-art, super-human performance (2) text-based question answering using bAbI tasks; and (3) complex reasoning about dynamic physical systems. using a curated dataset called Sort-of-CLEVR show that powerful convolutional networks do not have a general capacity to solve relational questions, but can gain this capacity when augmented with RNs. shows how a deep learning architecture equipped with an RN module can implicitly discover and learn to reason about entities and their relations.		2017
Anna Coenen 1 , Jonathan D. Nelson 2 , Todd M. Gureckis	Asking the right questions about human inquiry		2017
Ankit Kumar, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, Richard Socher	Ask Me Anything: Dynamic Memory Networks for Natural Language Processing	Memory Network	
Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel	Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources	Knowledge-based	
Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank	Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures		
Shayne Longpre, Sabeek Pradhan, Caiming Xiong, Richard Socher	**A Way Out of The Odyssey - Analyzing and Combining Recent Insights for LSTMs** **ABSTRACT** LSTMs have become a basic building block for many deep NLP models. In recent years, many improvements and variations have been proposed		2016

	for deep sequence models in general, and LSTMs in particular. We propose and analyze a series of augmentations and modifications to LSTM networks resulting in improved performance for text classification datasets. We observe compounding improvements on traditional LSTMs using Monte Carlo test-time model averaging, average pooling, and residual connections, along with four other suggested modifications. Our analysis provides a simple, reliable, and high quality baseline model.		
Yuke Zhu, Ce Zhang, Christopher Ré, Li Fei-Fei	**Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries** **Abstract:** The complexity of the visual world creates significant challenges for comprehensive visual understanding. In spite of recent successes in visual recognition, today's vision systems would still struggle to deal with visual queries that require a deeper reasoning. We propose a knowledge base (KB) framework to handle an assortment of visual queries, without the need to train new classifiers for new tasks. Building such a large-scale multimodal KB presents a major challenge of scalability. We cast a large-scale MRF into a KB representation, incorporating visual, textual and structured data, as well as their diverse relations. We introduce a scalable knowledge base construction system that is capable of building a KB with half billion variables and millions of parameters in a few hours. Our system achieves competitive results compared to purpose-built models on standard recognition and retrieval tasks, while exhibiting greater flexibility in answering richer visual queries.	Knowledge-based	2015
Kibeom Kim, Jin-Hwa Kim, Byoung-Tak Zhang	**Cambot: A Visual Conversation Robot for Interactive Engagement** **Abstract** —To achieve human-level artificial intelligence, it is crucial to develop algorithms which handle human-like visual and linguistic information. One of promising solutions is to use Multimodal Residual Networks (MRN) for the multimodal residual learning in assumption of visual question-answering tasks. It extends the idea of the deep residual learning, which learns joint representation from vision and language information effectively. While the MRN is handling with multidisciplinary problems of vision, language and integrated reasoning, a visual conversation robot can be a bridge to interact with humans. Cambot can be instantiated in any platform including robots, desktops and tablet PCs, which have a camera and microphone, engaging natural environmental situations of visual conversation for human interactions.		2016
Nigel G. Ward, David DeVault	**Challenges in Building Highly-Interactive Dialog Systems** **Abstract**		2016

	Spoken dialog researchers have recently demonstrated highly-interactive systems in several domains. This paper considers how to build on these advances to make systems more robust, easier to develop, and more scientifically significant. We identify key challenges whose solution would lead to improvements in dialog systems and beyond.		
Justin Johnson, Li Fei-Fei, Bharath Hariharan , C. Lawrence Zitnick, Laurens van der Maaten, Ross Girshick	CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning		
Aiwen Jiang, Fang Wang 2 Fatih Porikli 2 Yi Li ∗ 2,3	**Compositional Memory for Visual Question Answering** **Abstract:** Visual Question Answering (VQA) emerges as one of the most fascinating topics in computer vision recently. Many state of the art methods naively use holistic visual features with language features into a Long Short-Term Memory (LSTM) module, neglecting the sophisticated interaction between them. This coarse modeling also blocks the possibilities of exploring finer-grained local features that contribute to the question answering dynamically over time. This paper addresses this fundamental problem by directly modeling the temporal dynamics between language and all possible local image patches. When traversing the question words sequentially, our end-to-end approach explicitly fuses the features associated to the words and the ones available at multiple local patches in an attention mechanism, and further combines the fused information to generate dynamic messages, which we call episode. We then feed the episodes to a standard question answering module together with the contextual visual information and linguistic information. Motivated by recent practices in deep learning, we use auxiliary loss functions during training to improve the performance. Our experiments on two latest public datasets suggest that our method has a superior performance. Notably, on the DAQUAR dataset we advanced the state of the art by 6%, and we also evaluated our approach on the most recent MSCOCO-VQA dataset.	Memory Network	2015
Unnat Jain ∗ Ziyu Zhang ∗ Alexander Schwing	Creativity: Generating Diverse Questions using Variational Autoencoders		2017
Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, Devi Parikh	C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset		2017
Kyung-Min Kim 1,2 , Min-Oh Heo 1 , Seong-Ho Choi 1 , and Byoung-Tak Zhang	DeepStory: Video Story QA by Deep Embedded Memory Networks Abstract Question-answering (QA) on video contents is a significant challenge for achieving human-level intelligence as it involves both vision and language in real-world settings. Here we demonstrate the possibility of an AI agent performing video story QA by learning		2017

	from a large amount of cartoon videos. We develop a video-story learning model, i.e. Deep Embedded Memory Networks (DEMN), to reconstruct stories from a joint scene-dialogue video stream using a latent embedding space of observed data. The video stories are stored in a long-term memory component. For a given question, an LSTM-based attention model uses the long-term memory to recall the best question-story-answer triplet by focusing on specific words containing key information. We trained the DEMN on a novel QA dataset of children's cartoon video series, Pororo. The dataset contains 16,066 scene-dialogue pairs of 20.5-hour videos, 27,328 fine-grained sentences for scene description, and 8,913 story-related QA pairs. Our experimental results show that the DEMN outperforms other QA models. This is mainly due to 1) the reconstruction of video stories in a scene-dialogue combined form that utilize the latent embedding and 2) attention. DEMN also achieved state-of-the-art results on the MovieQA benchmark.		
Aishwarya Agrawal, Dhruv Batra, Devi Parikh, Aniruddha Kembhavi	Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering		
Huijuan Xu, Kate Saenko	Dual Attention Network for Visual Question Answering Abstract. Visual Question Answering (VQA) is a popular research problem that involves inferring answers to natural language questions about a given visual scene. Recent neural network approaches to VQA use attention to select relevant image features based on the question. In this paper, we propose a novel Dual Attention Network (DAN) that not only attends to image features, but also to question features. The selected linguistic and visual features are combined by a recurrent model to infer the final answer. We experiment with different question representations and do several ablation studies to evaluate the model on the challenging VQA dataset.	Attention	2016
Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, Tatsuya Harada	**DualNet: Domain-Invariant Network for Visual Question Answering** **Abstract:** VQA bridges the gap between images and language requires contents within the image understood as indicated by linguistic context of the question, to generate the accurate answers. critical to build an efficient embedding of images and texts. DualNet: takes advantage of discriminative power of both image and textual features by separately performing two operations. Building an ensemble of DualNet boosts the performance. effective in both real images and abstract scenes, in spite of significantly different properties of respective domain. outperform previous state-of-the-art methods in real images category without explicitly employing attention mechanism, outperformed our own state-of-the-art method in abstract scenes category, won the first place in VQA Challenge 2016.		2017

Caiming Xiong, Stephen Merity, Richard Socher	Dynamic Memory Networks for Visual and Textual Question Answering	Memory Network	
Youngjae Yu, Hyungjin Ko, Jongwook Choi, Gunhee Kim	End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering		2016
Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, Olivier Pietquin	**End-to-end optimization of goal-driven and visually grounded dialogue systems** **Abstract** End-to-end design of dialogue systems powerful tools such as encoder-decoder architectures for sequence-to-sequence learning. most approaches cast human-machine dialogue management as a supervised learning problem, aiming at predicting the next utterance of a participant given the full history of the dialogue. This vision is too simplistic to render the intrinsic planning problem inherent to dialogue as well as its grounded nature, making the context of a dialogue larger than the sole history. This is why only chit-chat and question answering tasks have been addressed so far using end-to-end architectures. introduce a **Deep Reinforcement Learning** method to optimize visually grounded task-oriented dialogues, based on the policy gradient algorithm. This approach is tested on a dataset of 120k dialogues collected through Mechanical Turk and provides encouraging results at solving both the problem of generating natural dialogues and the task of discovering a specific object in a complex picture.		
Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, Anthony Dick	Explicit Knowledge-based Reasoning for Visual Question Answering	Knowledge-based	2015
Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, Raffaella Bernardi	FOIL it! Find One mismatch between Image and Language caption		
Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, Anton van den Hengel	FVQA: Fact-based Visual Question Answering	Knowledge-based	2016
Ramakrishna Vedantam, Ian Fischer, Jonathan Huang	Generative Models of Visually Grounded Imagination		2017
Damien Teney, Lingqiao Liu,	**Graph-Structured Representations for Visual Question Answering**	structure	2017

Anton van den Hengel	**Abstract** improve VQA with structured representations of both scene contents and questions. A key challenge in VQA is to require joint reasoning over the visual and text domains. The predominant CNN/LSTM-based approach to VQA is limited by monolithic vector representations that largely ignore structure in the scene and in the question. CNN feature vectors cannot effectively capture situations as simple as multiple object instances, LSTMs process questions as series of words, which do not reflect the true complexity of language structure. propose to build graphs over the scene objects and over the question words, and we describe a deep neural network that exploits the structure in these representations. Achieves significant improvements over the state-of-the-art, increasing accuracy from 71.2% to 74.4% on the "abstract scenes" multiple-choice benchmark, and from 34.7% to 39.1% for the more challenging "balanced" scenes, i.e. image pairs with fine-grained differences and opposite yes/no answers to a same question.		
Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, Aaron Courville	GuessWhat?! Visual object discovery through multi-modal dialogue		
Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang	**HADAMARD PRODUCT FOR LOW-RANK BILINEAR POOLING** **ABSTRACT:** Bilinear models provide rich representations compared with linear models. applied in visual tasks as object recognition, segmentation, and visual question-answering, get state-of-the-art performances taking advantage of the expanded representations. bilinear representations tend to be high-dimensional, limiting the applicability to computationally complex tasks. propose low-rank bilinear pooling using Hadamard product for an efficient attention mechanism of multimodal learning. outperforms compact bilinear pooling in visual question-answering tasks the state-of-the-art results on the VQA dataset better parsimonious property.	Attention-Based	2017
iasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh	**Hierarchical Question-Image Co-Attention for Visual Question Answering** **Abstract** attention models for VQA generate spatial maps highlighting image regions relevant to answering the question. in addition to modeling "where to look" or visual attention, it is equally important to model "what words to listen to" or question attention. a novel co-attention model for VQA jointly reasons about image and question attention.	Attention	2016

	reasons about the question (and consequently the image via the co-attention mechanism) in a hierarchical fashion via a novel 1-dimensional CNN model. outperforms all reported methods, improving the state-of-the-art: （1）on the VQA dataset ：from 60.4% to 62.1%, （2）on the COCO-QA dataset：from 61.6% to 65.4%		
Arun Mallya, Svetlana Lazebnik	High-level Cues for Predicting Motivations		2017
Ferenc Huszár	HOW (NOT) TO TRAIN YOUR GENERATIVE MODEL: SCHEDULED SAMPLING, LIKELIHOOD, ADVERSARY?		
Justin Johnson, Judy Hoffman, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick	Inferring and Executing Programs for Visual Reasoning		2017
Sreyasi Nag Chowdhury, Niket Tandon, Gerhard Weikum	**Know2Look: Commonsense Knowledge for Visual Search** **Abstract** With the rise in popularity of social media, images accompanied by contextual text form a huge section of the web. However, search and retrieval of documents are still largely dependent on solely textual cues. Although visual cues have started to gain focus, the imperfection in object/scene detection do not lead to significantly improved results. We hypothesize that the use of background commonsense knowledge on query terms can significantly aid in retrieval of documents with associated images. To this end we deploy three different modalities - text, visual cues, and commonsense knowledge pertaining to the query - as a recipe for efficient search and retrieval.		
Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher,	Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning		
Yuke Zhu, Joseph J. Lim, Li Fei-Fei	**Knowledge Acquisition for Visual Question Answering via Iterative Querying** **Abstract** learn new skills and knowledge for problem solving. deal with open-ended questions in visual world. a neural-based approach acquiring task-driven information for VQA proposes queries acquire information from external data. Supporting evidence from sources: （1） human-curated （2） automatic encoded and stored into a memory bank. acquiring task-driven evidence improves performance on	Knowledge Acquisition Memory Bank Task-Driven Supported Evidence Iterative QA Iterative Querying Neural Based	2017

	Visual7W and VQA datasets; these queries offer certain level of interpretability in iterative QA model.		
Yun-Nung Chen, Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Jianfeng Gao, Li Deng	Knowledge as a Teacher: Knowledge-Guided Structural Attention Networks Abstract Natural language understanding (NLU) is a core component of a spoken dialogue system. RNN obtained strong results on NLU due to their superior ability of preserving sequential information over time. Traditionally, the NLU module tags semantic slots for utterances considering their flat structures, as the underlying RNN structure is a linear chain. However, natural language exhibits linguistic properties that provide rich, structured information for better understanding. This paper introduces a novel model, knowledge-guided structural attention networks (K-SAN), a generalization of RNN to additionally incorporate non-flat network topologies guided by prior knowledge. There are two characteristics: 1) important substructures can be captured from small training data, allowing the model to generalize to previously unseen test data; 2) the model automatically figures out the salient substructures that are essential to predict the semantic tags of the given sentences, so that the understanding performance can be improved. The experiments on the benchmark Air Travel Information System (ATIS) data show that the proposed K-SAN architecture can effectively extract salient knowledge from substructures with an attention mechanism, and outperform the performance of the state-of-the-art neural network based frameworks.	Knowledge	2016
Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, Dhruv Batra	Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning		
Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Kate Saenko	**Learning to Reason: End-to-End Module Networks for Visual Question Answering** **Abstract**: Natural language questions: (1) inherently compositional (2) easily answered by reasoning about their decomposition into modular sub-problems. to answer “is there an equal number of balls and boxes?” we can look for balls, look for boxes, count them, and compare the results. Neural Module Network (NMN) architecture [3,2] implements this approach to question answering by parsing questions into linguistic sub-structures assembling question-specific deep networks from smaller modules that each solve one subtask.		2017

	existing NMN implementations rely on brittle off-the-shelf parsers, restricted to the module configurations proposed by these parsers rather than learning them from data. propose End-to-End Module Networks (N2NMNs), learn to reason by directly predicting instance-specific network layouts without the aid of a parser. learns to generate network structures (by imitating expert demonstrations) while simultaneously learning network parameters (using the downstream task loss). results on CLEVR dataset targeted at compositional question answering show that N2NMNs achieve an error reduction of nearly 50% relative to state-of-the-art attentional approaches, discovering interpretable network architectures specialized for each question		
Peter D. Turney	**Leveraging Term Banks for Answering Complex Questions: A Case for Sparse Vectors** **Abstract** While open-domain QA systems proven effective for answering simple questions, they struggle with more complex questions. Our goal is to answer more complex questions reliably, without incurring a significant cost in knowledge resource construction to support the QA. One readily available knowledge resource is a term bank, enumerating the key concepts in a domain. We have developed an unsupervised learning approach that leverages a term bank to guide a QA system, by representing the terminological knowledge with thousands of specialized vector spaces. In experiments with complex science questions, we show that this approach significantly outperforms several state-of-the-art QA systems, demonstrating that significant leverage can be gained from continuous vector representations of domain terminology. In our experiments, we made the surprising discovery that dense, low-dimensional embeddings (used in many AI systems) were not the most effective representation, and that sparse, high-dimensional vector spaces performed better. We discuss the reasons for this, and the implications this may have for other projects that have assumed embeddings are the best continuous representation.		2017
Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, Min Sun	**Leveraging Video Descriptions to Learn Video Question Answering** **Abstract:** a scalable approach to learn video-based QA: answer free-form natural language questions about the contents of a video. automatically harvests a large number of videos and descriptions freely available online. a large number of candidate QA pairs automatically generated from descriptions rather than manually annotated. use these candidate QA pairs to train a number of video-based QA	Video QA	2016

	methods extended from: (1) MN (Sukhbaatar et al. 2015) (2) VQA (Antol et al. 2015) (3) SA (Yao et al. 2015) (4) SS (Venugopalan et al. 2015). to handle non-perfect candidate QA pairs, propose a self-paced learning procedure to iteratively identify them and mitigate their effects in training. evaluate performance on manually generated video-based QA pairs. self-paced learning procedure is effective, the extended SS model outperforms various baselines.		
Yash Goyal, Tejas Khot Douglas Summers-Stay, Dhruv Batra, Devi Parikh	Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering		2016
Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, Bohyung Han	**MarioQA: Answering Questions by Watching Gameplay Videos** **Abstract** a new benchmark dataset for VideoQA to evaluate algorithms' capability of spatio-temporal event understanding. Existing datasets: (1) either require very high-level reasoning from multi-modal information to find answers, (2) or is mostly composed of the questions that can be answered by watching a single frame. (3) not suitable to evaluate models' real capacity and flexibility for VideoQA. focus on event-centric questions that require understanding temporal relation between multiple events in videos. An interesting idea in dataset construction process is that question-answer pairs are automatically generated from Super Mario video game-plays given a set of question templates. tackle VideoQA problem in the new dataset, referred to as MarioQA, by proposing spatio-temporal attention models based on deep neural networks. the proposed deep neural network models with attention have meaningful performance improvement over several baselines.	VideoQA	2016
Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, Aaron Courville	**Modulating early visual processing by language** **Abstract** commonly assumed: language refers to high-level visual concepts while leaving low-level visual processing unaffected. This view dominates the current literature in computational models for language-vision tasks. visual and linguistic input processed independently before fused into a single representation. deviate from this classic pipeline: -- propose to modulate the entire visual processing by linguistic input. -- condition the batch normalization parameters of a pretrained residual		2017

	network (ResNet) on a language embedding. MOdulated RESnet (MORES), significantly improves strong baselines on two visual question answering tasks. modulating from the early stages of the visual processing is beneficial. ResNet image features are effectively grounded.		
Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, Sanja Fidler	**MovieQA: Understanding Stories in Movies through Question-Answering** **Abstract:** MovieQA dataset：evaluate automatic story comprehension from both video and text. 14,944 questions about 408 movies with high semantic diversity. questions range from simpler “Who” did “What” to “Whom”, to “Why” and “How” certain events occurred. Each question with five possible answers; a correct one and four deceiving answers by human annotators. contains multiple sources of information – video clips, plots, subtitles, scripts, and DVS [32]. analyze our data through various statistics and methods. extend existing QA techniques to show that question-answering with such open-ended semantics is hard.	Video QA	2016
Dongfei Yu, Jianlong Fu, Tao Mei, Yong Rui	**Multi-level Attention Networks for Visual Question Answering** **Abstract** VQA: automatically answer natural language questions with the reference to a given image. Compared with text-based QA, reasoning process on visual domain needs both effective semantic embedding and fine-grained visual understanding. Existing approaches predominantly infer answers from the abstract low-level visual features, while neglecting the modeling of high-level image semantics and the rich spatial context of regions. propose a multi-level attention network for visual question answering simultaneously reduce the semantic gap by semantic attention and benefit fine-grained spatial inference by visual attention. (1) generate semantic concepts from high-level semantics in CNN select question-related concepts as semantic attention. (2) encode region-based middle-level outputs from CNN into spatially-embedded representation by a bidirectional RNN, pinpoint the answer-related regions by multiple layer perceptron as visual attention. (3) jointly optimize semantic attention, visual attention and question embedding by a softmax classifier to infer the final answer. outperforms the-state-of-arts on two challenging VQA datasets.	Attention-based	2017
Akira Fukui* 1,2 Dong Huk Park* 1 Daylen Yang* 1	**Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding**	Attention-Based	2016

Anna Rohrbach* 1,3 Trevor Darrell 1 Marcus Rohrbach 1	**Abstract:** Modeling textual or visual information with vector representations trained from large language or visual datasets tasks as visual question answering require combining these vector representations Approaches to multimodal pooling include: (1) element-wise product or sum, (2) concatenation of the visual and textual representations. We hypothesize that these methods are not as expressive as an outer product of the visual and textual vectors. As the outer product is typically infeasible due to its high dimensionality, propose utilizing Multimodal Compact Bilinear pooling (MCB) to efficiently and expressively combine multi-modal features. We extensively evaluate MCB on the visual question answering and grounding tasks. We consistently show the benefit of MCB over ablations without MCB. For visual question answering, we present an architecture which uses MCB twice, once for predicting attention over spatial features and again to combine the attended representation with the question representation. This model outperforms the state-of-the-art on the Visual7W dataset and the VQA challenge.		
Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang	Multimodal Residual Learning for Visual QA Abstract Deep neural networks continue to advance the state-of-the-art of image recog- nition tasks with various methods. However, applications of these methods to multimodality remain limited. We present Multimodal Residual Networks (MRN) for the multimodal residual learning of visual question-answering, which extends the idea of the deep residual learning. Unlike the deep residual learning, MRN effectively learns the joint representation from vision and language information. The main idea is to use element-wise multiplication for the joint residual mappings exploiting the residual learning of the attentional models in recent studies. Var- ious alternative models introduced by multimodality are explored based on our study. We achieve the state-of-the-art results on the Visual QA dataset for both Open-Ended and Multiple-Choice tasks. Moreover, we introduce a novel method to visualize the attention effect of the joint representations for each learning block using back-propagation algorithm, even though the visual features are		

	collapsed without spatial information.		
Hedi Ben-younes Rémi Cadene Matthieu Cord Nicolas Thome	MUTAN: Multimodal Tucker Fusion for Visual Question Answering	Attention-Based	2017
Łukasz Kaiser Aidan N. Gomez Noam Shazeer Ashish Vaswani Niki Parmar Llion Jones Jakob Uszkoreit	One Model To Learn Them All		
Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Alan Yuille	**Recurrent Multimodal Interaction for Referring Image Segmentation** **Abstract** the problem of image segmentation given natural language descriptions, i.e. referring expressions. Existing works: （1）first modeling images and sentences independently and （2）then segment images by combining these two types of representations. We argue that learning word-to-image interaction is more native in the sense of jointly modeling two modalities for the image segmentation task, and we propose convolutional multimodal LSTM to encode the sequential interactions between individual words, visual information, and spatial information. We show that our proposed model outperforms the baseline model on benchmark datasets. In addition, we analyze the intermediate output of the proposed multimodal LSTM approach and empirically explains how this approach enforces a more effective word-to-image interaction.		2017
Allan Jabri Armand Joulin Laurens van der Maaten	Revisiting Visual Question Answering Baselines		
Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun	Rich Image Captioning in the Wild		
Andrei Barbu N.Siddharth Jeffrey Mark Siskind	Saying What You’re Looking For: Linguistics Meets Video Search		
Asli Celikyilmaz, Li Deng, Lihong Li, Chong Wang	**Scaffolding Networks: Incremental Learning and Teaching Through Questioning** **Abstract** a new paradigm of learning for: (1) reasoning (2) understanding (3) prediction the scaffolding network to implement this paradigm.	RL	2017

	The scaffolding network embodies an incremental learning approach formulated as a teacher-student network architecture to teach machines how to understand text and do reasoning. The key is the interactions between the teacher and the student through sequential questioning. The student observes each sentence in the text incrementally, it uses an attention-based neural net to discover and register the key information in relation to its current memory. Meanwhile, the teacher asks questions about the observed text, and the student network gets rewarded by correctly answering these questions. The entire network is updated continually using reinforcement learning. outperforms state-of-the-art methods. learns to do reasoning in a scalable way even with little human generated input.		
Yusuf Aytar, Carl Vondrick, Antonio Torralba	See, Hear, and Read: Deep Aligned Representations		
Zhe Gan † , Chuang Gan ∗ , Xiaodong He ‡ , Yunchen Pu † Kenneth Tran ‡ , Jianfeng Gao ‡ , Lawrence Carin † , Li Deng	Semantic Compositional Networks for Visual Captioning		
Bolei Zhou 1 , Yuandong Tian 2 , Sainbayar Sukhbaatar 2 , Arthur Szlam 2 , and Rob Fergus	Simple Baseline for Visual Question Answering		
Ted Zhang Dengxin Dai Tinne Tuytelaars	Speech-Based Visual Question Answering		
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola	**Stacked Attention Networks for Image Question Answering** **Abstract:** stacked attention networks (SANs) learn to answer natural language questions from images. use semantic representation of a question as query to search for the regions in an image related to the answer. image question answering (QA) often requires multiple steps of reasoning. develop a multiple-layer SAN in which we query an image multiple times to infer the answer progressively. significantly outperform previous state-of-the-art approaches on four image QA datasets. The visualization of the attention layers illustrates the progress that the SAN locates the relevant visual clues that lead to the answer of the question layer-by-layer.	Attention-Based	2016
Yuetan Lin Zhangyang Pang Donghui Wang ∗ Yueting Zhuang	Task-driven Visual Saliency and Attention-based Visual Question Answering		

Huan Ling 1 , Sanja Fidler	Teaching Machines to Describe Images via Natural Language Feedback		
Yunseok Jang 1 , Yale Song 2 , Youngjae Yu 1 , Youngjin Kim 1 , Gunhee Kim	TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering Abstract Vision and language understanding has emerged as a subject undergoing intense study in Artificial Intelligence. Among many tasks in this line of research, visual question answering (VQA) has been one of the most successful ones, where the goal is to learn a model that understands visual content at region-level details and finds their associations with pairs of questions and answers in the natural language form. Despite the rapid progress in the past few years, most existing work in VQA have focused primarily on images. In this paper, we focus on extending VQA to the video domain and contribute to the literature in three important ways. First, we propose three new tasks designed specifically for video VQA, which require spatio-temporal reasoning from videos to answer questions correctly. Next, we introduce a new large-scale dataset for video VQA named TGIF-QA that extends existing VQA work with our new tasks. Finally, we propose a dual-LSTM based approach with both spatial and temporal attention, and show its effectiveness over con- ventional VQA techniques through empirical evaluations.	Video-QA	2017
Andrew Shin, Yoshitaka Ushiku, Tatsuya Harada	The Color of the Cat is Gray: 1 Million Full-Sentences Visual Question Answering (FSVQA) Abstract Visual Question Answering (VQA) task has showcased a new stage of interaction between language and vi- sion, two of the most pivotal components of artificial intelligence. However, it has mostly focused on gener- ating short and repetitive answers, mostly single words, which fall short of rich linguistic capabilities of hu- mans. We introduce Full-Sentence Visual Question An- swering (FSVQA) dataset (www.mi.t.u-tokyo.ac. jp/static/projects/fsvqa), consisting of nearly 1 million pairs of questions and full-sentence answers for im- ages, built by applying a number of rule-based natural lan- guage processing techniques to original VQA dataset and captions in the MS COCO dataset. This poses many addi- tional complexities to conventional VQA task, and we pro- vide a baseline for approaching and evaluating the task, on top of which we invite the research community to build fur- ther improvements.		2017
Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra *,1 2	**The Promise of Premise: Harnessing Question Premises in Visual Question Answering** **Abstract**		

Dhruv Batra 2 Stefan Lee	an important observation: (1) questions about images often contain premises – objects and relationships implied by the question (2) reasoning about premises can help VQA models respond more intelligently to irrelevant or previously unseen questions. When presented with a question that is irrelevant to an image, state-of-the-art VQA models will still answer based purely on learned language biases, resulting in nonsensical or even misleading answers. We note that a visual question is irrelevant to an image if at least one of its premises is false (i.e. not depicted in the image). We leverage this observation to construct a dataset for Question Relevance Prediction and Explanation (QRPE) by searching for false premises. We train novel irrelevant question detection models and show that models that reason about premises consistently outperform models that do not. We also find that forcing standard VQA models to reason about premises during training can lead to improvements on tasks requiring compositional reasoning.		
Peng Wang * , Qi Wu * , Chunhua Shen, Anton van den Hengel	**The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions** Abstract One of the most intriguing features of the Visual Question Answering (VQA) challenge is the unpredictability of the questions. Extracting the information required to answer them demands a variety of image operations from detection and counting, to segmentation and reconstruction. To train a method to perform even one of these operations accurately from {image,question,answer} tuples would be challenging, but to aim to achieve them all with a limited set of such training data seems ambitious at best. We propose here instead a more general and scalable approach which exploits the fact that very good methods to achieve these operations already exist, and thus do not need to be trained. Our method thus learns how to exploit a set of external off-the-shelf algorithms to achieve its goal, an approach that has something in common with the Neural Turing Machine [10]. The core of our proposed method is a new co-attention model. In addition, the proposed approach generates human-readable reasons for its decision, and can still be trained end-to-end without ground truth reasons being given. We demonstrate the effectiveness on two publicly available datasets, Visual Genome and VQA, and show that it produces the state-of-the-art results in both cases.		2016
Issey Masuda Mora, Santiago Pascual de la Puente, Xavier Giro-i-Nieto	**Towards Automatic Generation of Question Answer Pairs from Images** **Abstract** generic field of Visual Question-Answering (VQA)		

	generate question-answer pairs based on an image. use the VQA dataset provided for the VQA challenge to train a Deep Neural Network which has the image as an input and two different outputs, the question and its associated answer.		
Hyeonwoo Noh, Bohyung Han	**Training Recurrent Answering Units with Joint Loss Minimization for VQA** **Abstract** a novel algorithm for visual question answering based on a RNN, every module in the network corresponds to a complete answering unit with attention mechanism by itself. The network optimized by minimizing loss aggregated from all the units share model parameters while receiving different information to compute attention probability. For training, model attends to a region within image feature map, updates its memory based on the question and attended image feature, and answers the question based on its memory state. This procedure is performed to compute loss in each step. multi-step inferences required to answer questions each problem may have a unique desirable number of steps, difficult to identify in practice. make the first unit in the network solve problems, but allow it to learn the knowledge from the rest of units by back-propagation unless it degrades the model. early-stop training each unit as soon as it starts to over-fit. Note that, since more complex models tend to over-fit on easier questions quickly, the last answering unit in the unfolded recurrent neural network is typically killed first while the first one remains last. We make a single-step prediction for a new question using the shared model. This strategy works better than the other options within our framework since the selected model is trained effectively from all units without overfitting. achieves the state-of-the-art performance on the standard benchmark dataset without data augmentation.	Attention-Based	2016
Linchao Zhu § Zhongwen Xu † Yi Yang † Alexander G. Hauptmann	**Uncovering Temporal Context for Video Question and Answering** **Abstract:** in temporal domain to: (1) infer the past (2) describe the present (3) predict the future **RNN encoder-decoder** to learn temporal structures of videos a **dual-channel ranking loss** to answer multiple-choice questions. explore approaches for finer understanding of video content using question form of “fill-in-the-blank” collect 109,895 video clips over 1,000 hours from TACoS, MPII-MD, MEDTest 14 datasets, 390,744 questions generated from annotations. significantly outperforms the baselines	Video QA	2015
Marc Bolaños	**VIBIKNet: Visual Bidirectional Kernelized Network for Visual**		2016

1,2 , Álvaro Peris 3 , Francisco Casacuberta 3 , Petia Radeva	**Question Answering** **Abstract** VIBIKNet integrating Kernelized CNN and LSTM units to generate an answer given a question about an image. an optimal trade-off between accuracy and computational load, in terms of memory and time consumption. validate our method on the VQA challenge dataset and compare it to the top performing methods in order to illustrate its performance and speed.		
Amir Mazaheri, Dong Zhang, Mubarak Shah	**Video Fill In the Blank using LR/RL LSTMs with Spatial-Temporal Attentions** Abstract Given a video and a description sentence with one missing word (we call it the "source sentence"), Video-Fill-In-the-Blank (VFIB) problem: find the missing word automatically. The contextual information of the sentence, as well as visual cues from the video, are important to infer the missing word accurately. Since the source sentence is broken into two fragments: the sentence's left fragment (before the blank) and the sentence's right fragment (after the blank), traditional Recurrent Neural Networks cannot encode this structure accurately because of many possible variations of the missing word in terms of the location and type of the word in the source sentence. For example, a missing word can be the first word or be in the middle of the sentence and it can be a verb or an adjective. propose a framework to tackle the textual encoding: Two separate LSTMs (the LR and RL LSTMs) are employed to encode the left and right sentence fragments and a novel structure is introduced to combine each fragment with an external memory corresponding the opposite fragments. For the visual encoding, end-to-end spatial and temporal attention models are employed to select discriminative visual representations to find the missing word. In the experiments, we demonstrate the superior performance of the proposed method on challenging VFIB problem. Furthermore, we introduce an extended and more generalized version of VFIB, which is not limited to a single blank. Our experiments indicate the generalization capability of our method in dealing with such more realistic scenarios.		
Fereshteh Sadeghi, Santosh K. Divvala, Ali Farhadi	**VisKE: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases** **Abstract** How can we know whether a statement about our world is valid. For example, given a relationship between a pair of entities e.g., 'eat(horse, hay)', how can we know whether this relationship is true or false in general. Gathering such knowledge about entities and their relationships is one of	Knowledge Extraction	2015

	the fundamental challenges in knowledge extraction. Most previous works on knowledge extraction have focused purely on text-driven reasoning for verifying relation phrases. we introduce the problem of visual verification of relation phrases developed a Visual Knowledge Extraction system called VisKE. Given a verb-based relation phrase between common nouns, assess its validity by jointly analyzing over text and images reasoning about the spatial consistency of the relative configurations of the entities and the relation involved. involves no explicit human supervision enabling large-scale analysis. verified over 12000 relation phrases. enrich textual knowledge bases by improving their recall augment open-domain question-answer reasoning.		
Abhishek Das, Satwik Kottur, Khushi Gupta 2 * , Avi Singh 3 * , Deshraj Yadav 1 , José M.F. Moura 2 , Devi Parikh 1 , Dhruv Batra	Visual Dialog		2017
Nicholas Watters, Andrea Tacchetti, Théophane Weber, Razvan Pascanu, Peter Battaglia, and Daniel Zoran	Visual Interaction Networks		2017
Qi Wu, Damien Teney, Peng Wang, Chunhua Shen * , Anthony Dick, Anton van den Hengel	Visual Question Answering: A Survey of Methods and Datasets		2016
Kushal Kafle Christopher Kanan	Visual Question Answering: Datasets, Algorithms, and Future Challenges		2017
Ruiyu Li, Jiaya Jia	**Visual Question Answering with Question Representation Update (QRU)** **Abstract** reasoning over natural language questions and visual images. Given a natural language question about an image, our model updates the question representation iteratively by selecting image regions relevant to the query and learns to give the correct answer. Our model contains several reasoning layers, exploiting complex visual relations in the visual question answering (VQA) task. The proposed network is end-to-end trainable through back-propagation, where its weights are initialized using pre-trained CNN and GRU. Our method is evaluated on challenging datasets of COCO-QA [19] and VQA [2] and yields state-of-the-art performance.		2016
Ting-Hao (Kenneth) Huang 1* , Francis	Visual Storytelling		2016

Ferraro 2∗ , Nasrin Mostafazadeh 3 , Ishan Misra 1 , Aishwarya Agrawal 4 , Jacob Devlin 6 , Ross Girshick 5 , Xiaodong He 6 , Pushmeet Kohli 6 , Dhruv Batra 4 , C. Lawrence Zitnick 5 , Devi Parikh 4 , Lucy Vanderwende 6 , Michel Galley 6 , Margaret Mitchell 6			
Jeffrey P. Bigham Chandrika Jayant Hanjie Ji † , Greg Little § , Andrew Miller ɣ , Robert C. Miller § , Robin Miller † , Aubrey Tatarowicz § , Brandyn White ‡ , Samuel White † , and Tom Yeh	VizWiz: Nearly Real-time Answers to Visual Questions		2010
Rene Grzeszick, Gernot A. Fink	**Zero-shot object prediction using semantic scene knowledge** **Abstract** the semantic relations between scenes and objects for visual object recognition. Semantic knowledge: a powerful source of information especially in scenarios with few or no annotated training samples. zero-shot or few-shot recognition and often build on visual attributes. Here, instead of relying on various visual attributes, a more direct way is pursued: after recognizing the scene that is depicted in an image, semantic relations between scenes and objects are used for predicting the presence of objects in an unsupervised manner. Most importantly, relations between scenes and objects can easily be obtained from external sources such as large scale text corpora from the web and, therefore, do not require tremendous manual labeling efforts. It will be shown that in cluttered scenes, where visual recognition is difficult, scene knowledge is an important cue for predicting objects.	knowledge	2016
Damien Teney Anton van den Hengel	Zero-Shot Visual Question Answering		2016