Xiaodong He and Li Deng

Deep Learning for Image-to-Text Generation

A technical overview

©ISTOCKPHOTO.COM/ZAPP2PHOTO

Digital Object Identifier 10.1109/MSP.2017.2741510
Date of publication: 13 November 2017

Generating a natural language description from an image is an emerging interdisciplinary problem at the intersection of computer vision, natural language processing, and artificial intelligence (AI). This task, often referred to as *image* or *visual captioning*, forms the technical foundation of many important applications, such as semantic visual search, visual intelligence in chatting robots, photo and video sharing in social media, and aid for visually impaired people to perceive surrounding visual content. Thanks to the recent advances in deep learning, the AI research community has witnessed tremendous progress in visual captioning in recent years. In this article, we will first summarize this exciting emerging visual captioning area. We will then analyze the key development and the major progress the community has made, their impact in both research and industry deployment, and what lies ahead in future breakthroughs.

Introduction

It has been long envisioned that machines one day will understand the visual world at a human level of intelligence. Thanks to the progress in deep learning [15], [36], [59], [60], [69], researchers can now build very deep convolutional neural networks (CNNs) and achieve an impressively low error rate for tasks like large-scale image classification [9], [15], [23]. In these tasks, one way for researchers to train a model to predict the category of a given image is to first annotate each image in a training set with a label from a predefined set of categories. Through such fully supervised training, the computer learns how to classify an image.

However, in tasks like image classification, the content of an image is usually simple, containing a predominant object to be classified. The situation could be much more challenging when we want computers to understand complex scenes. Image captioning is one such task. The challenges come from two perspectives. First, to generate a semantically meaningful and syntactically fluent caption, the system needs to detect salient semantic concepts in the image, understand the relationships among them, and compose a coherent description about the overall content of the image, which involves language and common-sense knowledge

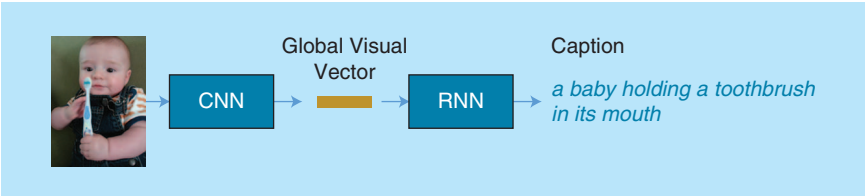

FIGURE 1. An illustration of the CNN-RNN-based image captioning framework.

modeling beyond object recognition. In addition, due to the complexity of scenes in the image, it is difficult to represent all fine-grained, subtle differences among them with the simple attribute of category. The supervision for training image captioning models is a full description of the content of the image in natural language, which is sometimes ambiguous and lacks fine-grained alignments between the subregions in the image and the words in the description.

Moreover, unlike image classification tasks, where we can easily tell if the classification output is correct or wrong after comparing it to the ground truth, there are multiple valid ways to describe the content of an image. It is not easy to tell if the generated caption is correct or not, at what degree. In practice, human studies are often employed to judge the quality of the caption given an image. However, since human evaluation is costly and time-consuming, many automatic metrics are proposed, which could serve as a proxy mainly for speeding up the development cycle of the system.

Early approaches to image captioning can be roughly divided into two families. The first one is based on template matching [6], [16], [17]. These approaches start from detecting objects, actions, scenes, and attributes in images and then fill them into a hand-designed and rigid sentence template. The captions generated by these approaches are not always fluent and expressive. The second family is grounded on retrieval-based approaches, which first select a set of the visually similar images from a large database and then transfer the captions of retrieved images to fit the query image [10], [20]. There is little flexibility to modify words based on the content of the query image, since they directly rely on captions of training images and cannot generate new captions.

Deep neural networks can potentially address both of these issues by generating fluent and expressive captions, which can also generalize beyond those in the train set. In particular, recent successes of using neural networks in image classification [9], [15], [23] and object detection [8] have motivated strong interest in using neural networks for visual captioning.

Major deep-learning paradigms for image captioning

The end-to-end framework

Vector-to-sequence learning

Motivated by the recent success of sequence-to-sequence learning in machine translation [37]–[39], researchers studied an end-to-end encoder-decoder framework for image captioning [2]–[4], [12], [26]. Figure 1 illustrates a typical encoder-decoder-based captioning system [26].

In such a framework, first the raw image is encoded by a global visual feature vector which represents the overall semantic information of the image, via deep CNNs. As illustrated in Figure 2, a CNN consists of several convolutional, max-pooling, response-normalization, and fully connected layers. This architecture has been very successful for large-scale image classification [21], and the learned features have shown to transfer to a broad variety of vision tasks [40]. Usually, given a raw image, the activation

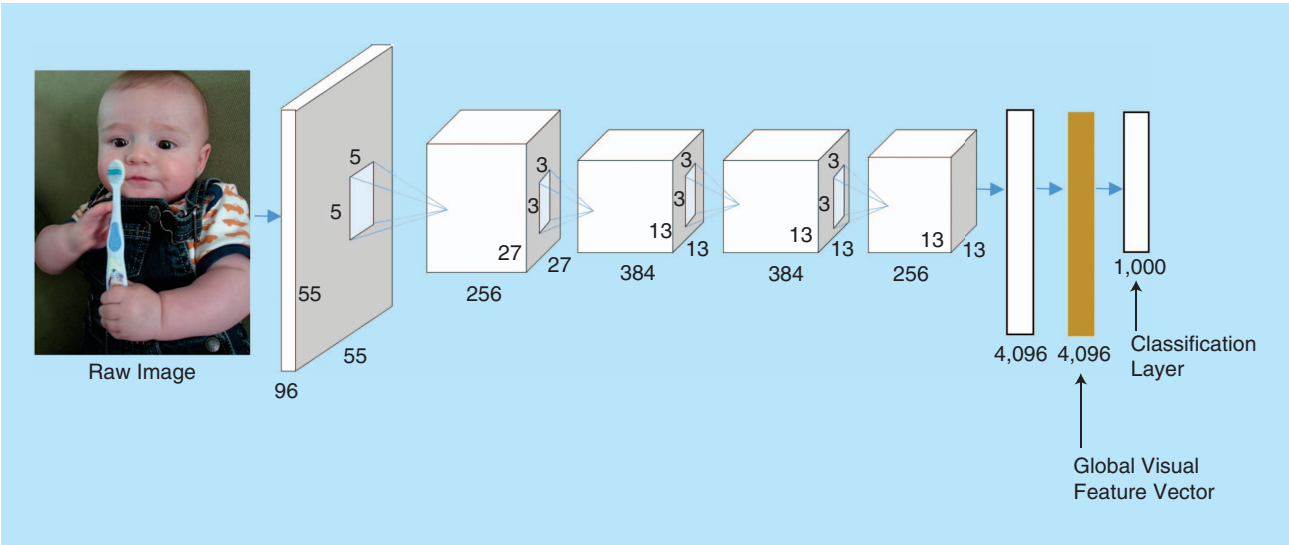

FIGURE 2. An illustration of a deep CNN such as the AlexNet [15]. The CNN is trained for a 1,000-class image classification task on the large-scale ImageNet data set [41]. The last layer of the AlexNet contains 1,000 nodes, each corresponding to a category. The second last fully connected dense layer is usually extracted as the global visual feature vector, representing the semantic content of the overall images.

values at the second last fully connected layer are extracted as the global visual feature vector.

Once the global visual vector is extracted, it is then fed into a recurrent neural network (RNN)-based decoder for caption generation, as illustrated in Figure 3. In practice, a long-short memory network (LSTM) [40] or gated recurrent unit (GRU) [39] variation of the RNN is often used; both have been shown to be more efficient and effective in training and capturing long-span language dependencies than vanilla RNNs [38], [39], and both have found successful applications in action recognition tasks [62], [63].

The representative set of studies using the aforementioned end-to-end framework include [2]–[4], [7], [11]–[13], [19], and [26] for image captioning and [1], [21] [24], [25], and [32] for video captioning. The differences of the various methods mainly lie in the types of CNN architectures and the RNN-based language models. For example, the vanilla RNN was used in [12] and [19], while the LSTM was used in [26]. The visual feature vector was only fed into the RNN once at the first time step in [26], while it was used at each time step of the RNN in [19].

The attention mechanism

Most recently, [29] utilized an attention-based mechanism to learn where to focus in the image during caption generation. The attention architecture is illustrated in Figure 4. Different from the simple encoder-decoder approach, the attention-based approach first uses the CNN to not only generate a global visual vector but also generate a set of visual vectors for subregions in the image. These subregion vectors can be extracted from a lower convolutional layer in the CNN. Then, in language generation, at each step of generating a new word, the RNN will refer to these subregion vectors and determine the likelihood that each of the subregions is relevant to the current state to generate the word. Eventually, the attention mechanism will form a contextual vector, which is a sum of subregional visual vectors weighted by the likelihood of relevance, for the RNN to decode the next new word.

This work was followed by [30], which introduced a "review" module to improve the attention mechanism and further by [18], which proposed a method to improve the correctness of visual attention. More recently, based on object detection, a bottom-up attention model was proposed in [64], which demonstrated a state-of-the-art performance on image captioning. In the end-to-end framework, all of the model parameters, including the CNN, the RNN, and the attention model, are trained jointly in an end-to-end fashion; hence, the term *end to end*.

A compositional framework

Different from the end-to-end encoder-decoder framework previously described, a separate class of image-to-text approaches uses an explicit semantic-concept-detection process for caption generation. The detection model and other modules are often trained separately. Figure 5 illustrates a semantic-concept-detection-based compositional approach proposed by Fang et al. [5].

In this framework, the first step in the caption generation pipeline detects a set of semantic concepts, known as *tags* or *attributes*, that are likely to be part of the image's description. These tags may belong to any part of speech, including nouns, verbs, and adjectives. Unlike image classification, standard supervised learning techniques are not directly applicable for learning detectors since the supervision only contains the whole image and the human-annotated whole sentence of caption, while the image bounding boxes corresponding to the words are unknown. To address this issue, [5] proposed learning the detectors using the weakly supervised approach of multiple instance learning (MIL) [42], [43], while in [33], this problem is treated as a multilabel classification task.

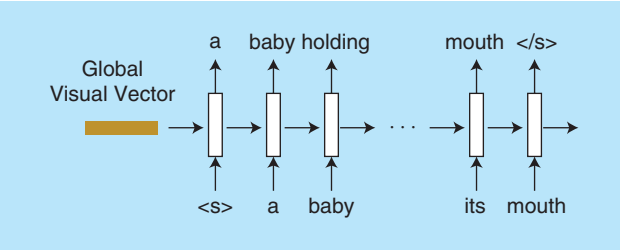

FIGURE 3. An illustration of an RNN-based caption decoder. At the initial step, the global visual vector, which represents the overall semantic meaning of the image, is fed into the RNN to compute the hidden layer at the first step while the sentence-start symbol <s> is used as the input to the hidden layer at the first step. Then the first word is generated from the hidden layer. Continuing this process, the word generated in the previous step becomes the input to the hidden layer at the next step to generate the next word. This generation process keeps going until the sentence-end symbol, </s>, is generated.

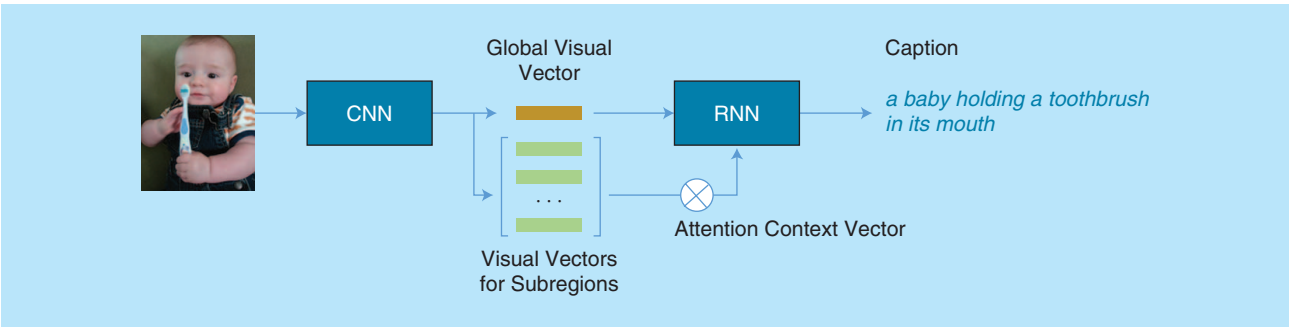

FIGURE 4. An illustration of the attention mechanism in the image caption generation process.

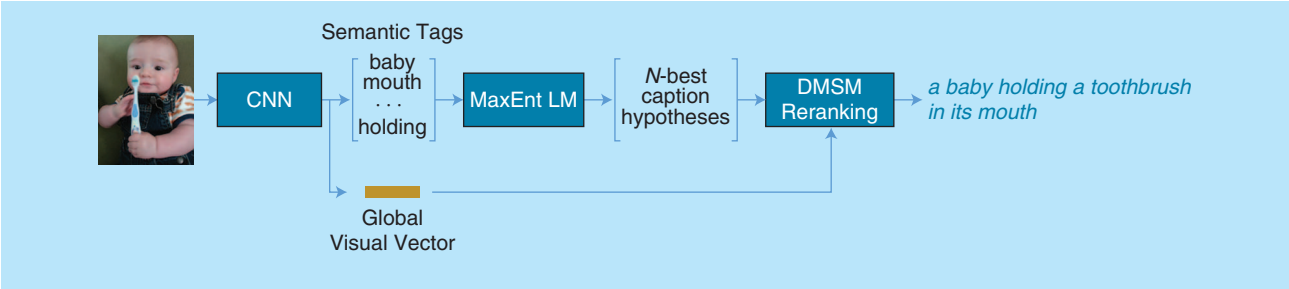

FIGURE 5. An illustration of a semantic-concept-detection-based compositional approach [5].

In [5], the detected tags are then fed into an n-gram-based max-entropy language model to generate a list of caption hypotheses. Each hypothesis is a full sentence that covers certain tags and is regularized by the syntax modeled by the language model, which defines the probability distribution over word sequences.

All of these hypotheses are then reranked by a linear combination of features computed over an entire sentence and the whole image, including sentence length, language model scores, and semantic similarity between the overall image and an entire caption hypothesis. Among them, the image-caption semantic similarity is computed by a deep multimodal similarity model (DMSM), which consists of a pair of neural networks, one for mapping each input modality, image, and language, to be vectors in a common semantic space. Image-caption semantic similarity is then defined as the cosine similarity between their vectors.

Compared to the end-to-end framework, the compositional approach provides better flexibility and scalability in system development and deployment and facilitates exploiting various data sources to optimizing the performance of different modules more effectively, rather than learn all of the models on limited image-caption paired data. On the other hand, the end-to-end model usually has a simpler architecture and can optimize the overall system jointly for a better performance.

More recently, a class of models has been proposed to integrate explicit semantic-concept detection in an encoder-decoder framework. A general diagram of this class of models is illustrated in Figure 6. For example, [1] applied retrieved sentences as additional semantic information to guide the LSTM when generating captions, while [31] and [33] applied a semantic-concept-detection process before generating sentences. In [7], a semantic compositional network is constructed based on the probability of detected semantic concepts for composing captions.

Other related work

Other related work also learns a joint embedding of visual features and associated captions, including [5] for image captioning and [21] for video captioning. Most recently, [27] has looked into generating dense image captions for individual regions in images. In addition, a variational autoencoder was developed in [22] for image captioning. Also motivated by its recent success, researchers proposed a set of reinforcement learning-based algorithms to directly optimize the model for specific rewards. For example, [67] proposed a self-critical sequence training algorithm. It uses the REINFORCE algorithm to optimize a particular evaluation metric that is usually not differentiable and therefore not easy to optimize by conventional gradient-based methods. In [65], within the actor-critic framework, a policy network and a value network are learned to generate the caption by optimizing a visual semantic reward, which measures the similarity between the image and generated caption. Relevant to image-caption generation, models based on generative adversarial networks (GANs) recently have been proposed for text generation. Among them, SeqGAN [68] models the generator as a stochastic policy in reinforcement learning for discrete outputs like texts, while RankGAN [66] proposed a ranking-based loss for the discriminator, which gives better assessment of the quality of the generated text and therefore leads to a better generator.

Metrics

The quality of the automatically generated captions is evaluated and reported in the literature in both automatic metrics and human studies. Commonly used automatic metrics include BLEU [45], METEOR [44], CIDEr [46], and SPICE [47]. BLEU [45] is widely used in machine translation and measures the fraction of n-grams (up to four grams) that are in common between a hypothesis and a reference or set of references. METEOR [44] instead measures unigram precision and recall, but extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens. CIDEr [46] also measures the n-gram match between the caption hypothesis and the references, while the n-grams are weighted by term frequency–inverse document frequency (TF-IDF). On the other hand, SPICE [47] measures the F1 score of semantic propositional content contained in image captions given the references, and therefore it has the best correlation to human judgment [47]. These automatic metrics can be computed efficiently. They can greatly speed up the development of image captioning algorithms. However, all of these automatic metrics are known to only roughly correlate with human judgment [50].

Benchmarks

Researchers have created many data sets to facilitate the research of image captioning. The Flickr data set [49] and the PASCAL sentence data set [48] were created for facilitating

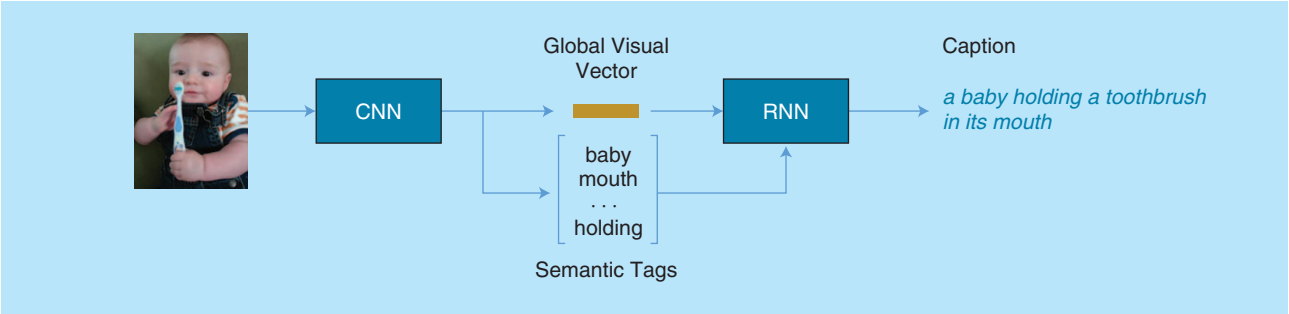

FIGURE 6. An illustration of integrate explicit semantic-concept-detection in an encoder-decoder framework.

the research of image captioning. More recently, Microsoft sponsored the creation of the Common Objects in Context (COCO) data set [51], the largest image captioning data set available to the public today. The availability of the large-scale data sets significantly prompted research in image captioning in the last several years.

In 2015, approximately 15 groups participated in the COCO Captioning Challenge [52]. The entries in the challenge are evaluated by human judgment. Five human judge metrics are listed in Table 1. In the competition, all entries are assessed based on the results from metric 1 (M1) and metric 2 (M2).The other metrics have been used as diagnostic and interpretation of the results. Specifically, in evaluation, each task presents a human judge with an image and two captions: one is automatically generated, and the other is a human caption. For M1, the judge is asked to select which caption better describes the image, or to choose the "same" option when they are of equal quality. For M2, the judge is asked to tell which of the two captions is generated by a human. If the judge chooses the automatically generated caption, or chooses the "cannot tell" option, it is considered to have passed the Turing test. Table 2 tabulates the results of the 15 entries in the 2015 COCO Captioning Challenge. Among them, the Microsoft Research entry (MSR) achieves the best performance on the Turing test metric, while the Google team outperforms others in the percentage of captions that were as good or better than human captions. Overall, Microsoft Research and Google jointly received first prize in the 2015 COCO Image Captioning Challenge. The results of two special systems, human and random, are also included for reference.

Microsoft Research and Google jointly received first prize in the 2015 COCO Image Captioning Challenge.

There are more systems that have been developed since the 2015 COCO competition. However, due to the high cost, human judging was no longer performed. Instead, the organizers of the COCO benchmark set up an automatic evaluation server. The server can receive the captions generated by a new system and then evaluate and report the results on the blind test set in automatic metrics. Table 3 summarizes the top 24 entries plus the human system as of August 2017, ranked by SPICE, using 40 references per image [52]. Note that these 24 systems outperform the human system in all automatic metrics except SPICE. However, in human judgment, it is likely that the human system still has a lead, given that in Table 2 there is a huge gap between the best systems and a human.

Table 1. Human evaluation metrics in the 2015 COCO Captioning Challenge.

Metric	Comment
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing test.
M3	Average correctness of the captions on a scale from one to five (incorrect–correct).
M4	Average amount of detail of the captions on a scale from one to five (lack of details–very detailed).
M5	Percentage of captions that are similar to human description.

Industrial deployment

Given the fast progress in the research community, the industry started deploying image captioning services. In March 2016, Microsoft released the first public image captioning application programming interface as a cloud service [53]. To showcase the usage of the functionality, it deployed a web application called CaptionBot (http://CaptionBot.ai) which captions arbitrary pictures users uploaded [33]. The service also supports applications like Seeing AI, designed for the low-vision community, that narrate the world around people who are blind or visually impaired [71]. More recently, Microsoft further deployed the caption service in its widely used product Office, specifically, Word and PowerPoint, for automatically generating alt-text, i.e., text descriptions of pictures, for accessibility [61]. Facebook released an automatic image captioning tool that provides a list of objects and scenes identified in a photo [34]. Meanwhile, although the service has not yet been deployed publicly, Google open sourced its image captioning system for the community [35]. With all of these industrial-scale deployment and open-source projects, a massive number of images and user feedback in real-world scenarios are collected and serve as training data to continuously

Table 2. Human evaluation results of entries in the 2015 COCO Captioning Challenge.

Entry	M1	M2	M3	M4	M5	Date
Human	0.638	0.675	4.836	3.428	0.352	23 March 2015
Google	0.273	0.317	4.107	2.742	0.233	29 May 2015
MSR	0.268	0.322	4.137	2.662	0.234	8 April 2015
Montreal/Toronto	0.262	0.272	3.932	2.832	0.197	14 May 2015
MSR Captivator	0.25	0.301	4.149	2.565	0.233	28 May 2015
Berkeley LRCN	0.246	0.268	3.924	2.786	0.204	25 April 2015
m-RNN	0.223	0.252	3.897	2.595	0.202	30 May 2015
Nearest Neighbor	0.216	0.255	3.801	2.716	0.196	15 May 2015
PicSOM	0.202	0.25	3.965	2.552	0.182	26 May 2015
Brno University	0.194	0.213	3.079	3.482	0.154	29 May 2015
m-RNN (Baidu/UCLA)	0.19	0.241	3.831	2.548	0.195	26 May 2015
MIL	0.168	0.197	3.349	2.915	0.159	29 May 2015
MLBL	0.167	0.196	3.659	2.42	0.156	10 April 2015
NeuralTalk	0.166	0.192	3.436	2.742	0.147	15 April 2015
ACVT	0.154	0.19	3.516	2.599	0.155	26 May 2015
Tsinghua Bigeye	0.1	0.146	3.51	2.163	0.116	23 April 2015
Random	0.007	0.02	1.084	3.247	0.013	29 May 2015

Table 3. The state-of-the-art image captioning systems in automatic metrics (as of 8 December 2016).

Entry	CIDEr-D	METEOR	BLEU-4	SPICE (x10)	Date
Watson Multimodal	1.123	0.268	0.344	0.204	16 November 2016
DONOT_FAIL_AGAIN	1.01	0.262	0.32	0.199	22 November 2016
Human	0.854	0.252	0.217	0.198	23 March 2015
MSM@MSRA	1.049	0.266	0.343	0.197	25 October 2016
MetaMind/VT_GT	1.042	0.264	0.336	0.197	1 December 2016
ATT-IMG (MSM@MSRA)	1.023	0.262	0.34	0.193	13 June 2016
G-RMI(PG-SPIDEr-TAG)	1.042	0.255	0.331	0.192	11 November 2016
DLTC@MSR	1.003	0.257	0.331	0.19	4 September 2016
Postech_CV	0.987	0.255	0.321	0.19	13 June 2016
G-RMI (PG-BCMR)	1.013	0.257	0.332	0.187	30 October 2016
feng	0.986	0.255	0.323	0.187	6 November 2016
THU_MIG	0.969	0.251	0.323	0.186	3 June 2016
MSR	0.912	0.247	0.291	0.186	8 April 2015
reviewnet	0.965	0.256	0.313	0.185	24 October 2016
Dalab_Master_Thesis	0.96	0.253	0.316	0.183	28 November 2016
ChalLS	0.955	0.252	0.309	0.183	21 May 2016
ATT_VC_REG	0.964	0.254	0.317	0.182	3 December 2016
AugmentCNNwithDe	0.956	0.251	0.315	0.182	29 March 2016
AT	0.943	0.25	0.316	0.182	29 October 2015
Google	0.943	0.254	0.309	0.182	29 May 2015
TsinghuaBigeye	0.939	0.248	0.314	0.181	9 May 2016

improve the performance of the system and stimulate new researches in deep visual understanding.

Outlook

Image-to-text generation is an important interdisciplinary area across computer vision and natural language processing. It also forms the technical foundation of many important applications. Thanks to deep-learning technologies, we have seen significant progress in this area in recent years. In this article, we have reviewed the key developments that the community has made and their impact in both research and industry deployment. Looking forward, image captioning will be a key subarea in the image–natural language multimodal intelligence field. A number of new problems in this field have been proposed lately, including visual question answering [54], [55], [70], visual storytelling [58], visually grounded dialog [56], and image synthesis from text description [57], [72]. The progress in multimodal intelligence is critical for building more general AI abilities in the future, and we hope the overview provided in this article can encourage students and researchers to enter and contribute to this exciting AI area.

Authors

Xiaodong He (xiaohe@microsoft.com) received his bachelor's degree from Tsinghua University, Beijing, China, in 1996, his M.S. degree from the Chinese Academy of Sciences, Beijing, in 1999, and his Ph.D. degree from the University of Missouri–Columbia in 2003. He is a principal researcher in the Deep Learning Group of Microsoft Research, Redmond, Washington. He is also an affiliate professor in the Department of Electrical Engineering and Computer Engineering at the University of Washington, Seattle. His research interests are mainly in artificial intelligence areas including deep learning, natural language processing, computer vision, speech, information retrieval, and knowledge representation. He received several awards including the Outstanding Paper Award at the 2015 Conference of the Association for Computational Linguistics (ACL). He has held editorial positions on several IEEE journals, was the area chair for the North American Chapter of the 2015 Conference of the ACL, and served on the organizing committee/program committee of major speech and language processing conferences. He is a Senior Member of the IEEE.

Li Deng (l.deng@ieee.org) received the Ph.D. degree from the University of Wisconsin-Madison in 1987. He was an assistant professor (1989–1992), tenured associate professor (1992–1996), and full professor (1996–1999) at the University of Waterloo, Ontario, Canada. In 1999, he joined Microsoft Research, Redmond, Washington, where he currently leads the research and development of deep learning as a partner research manager of its Deep Learning Technology Center, and where he is a chief scientist of artificial intelligence. Since 2000, he has also been an affiliate full professor and graduate committee member at the University of Washington, Seattle. He is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association. He served on the Board of Governors of the IEEE Signal Processing Society (SPS) (2008–2010), and as editor-in-chief of *IEEE Signal Processing Magazine* (2009–2011), which earned the highest impact factor in 2010 and 2011 among all IEEE publications and for which he received the 2012 IEEE SPS Meritorious Service Award. He recently joined Citadel as its chief artificial intelligence officer.

References

[1] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. Int. Conf. Learning Representations*, 2016.

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

[3] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," in *roc. 53rd Annu. Meeting Association Computational Linguistics and the 7th Int. Joint Conf. Natural Language Processing,* 2015, Beijing, China, pp. 100–105.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Doll´ar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. "From captions to visual concepts and back," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, Every picture tells a story: Generating sentences from images. in *Proc. European Conf. Computer Vision*, 2010.

[7] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[8] R. Girshick. "Fast r-CNN," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 1440–1448.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp 770–778.

[10] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013.

[11] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 2407–2415.

[12] Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[13] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Multimodal neural language models," in *Proc. Int. Conf. Machine Learning*, 2014.

[14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv Preprint*, arXiv:1602.07332, 2016.

[15] Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Conf. Neural Information Processing Systems*, 2012.

[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. "Babytalk: Understanding and generating simple image descriptions," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1601–1608.

[17] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Computational Natural Language Learning*, 2011, pp. 220–228.

[18] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," *arXiv Preprint*, arXiv:1605.09553, 2016.

[19] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learning Representations*, 2015.

[20] V. Ordonez, G. Kulkarni, T. L. Berg, V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Conf. Neural Information Processing Systems*, 2011.

[21] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4594–4602.

[22] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Proc. Conf. Neural Information Processing Systems*, 2016.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Comput. Sci. Conf.*, 2014.

[24] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North American Chapter Association Computational Linguistics: Human Language Technologies*, 2015, pp. 1494–1505.

[25] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 4534–4542.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and tell: A neural image caption generator," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[27] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4565–4574.

[28] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v d. Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 203–212.

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

[30] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen, "Review networks for caption generation," in *Proc. Conf. Neural Information Processing Systems*, 2016.

[31] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[32] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4584–4593.

[33] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and, and C. Sienkiewicz, "Rich image captioning in the wild. Deep Vision Workshop," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 434–441.

[34] S. Wu, J. Wieland, O. Farivar, and J. Schiller. "Automatic Alt-text: Computer-generated image descriptions for blind users on a social network service," in *Proc. 20th ACM Conf. Computer Supported Cooperative Work and Social Computing*, 2017.

[35] C. Shallue. (2016). Open-source code on show and tell: A neural image caption generator. [Online]. Available: https://github.com/tensorflow/models/tree/master/im2txt

[36] L. Deng and D. Yu, *Deep Learning: Methods and Applications*, NOW Publishers, 2014.

[37] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Conf. Neural Information Processing Systems*, 2014.

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representations*, 2015.

[39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Machine Learning*, 2015.

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 98, pp. 1735–1780 1997.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[42] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," in *Proc. Int. Conf. Machine Learning*, 2014.

[43] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Proc. Conf. Neural Information Processing Systems*, 2005.

[44] S. Banerjee and A. Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Association Computational Linguistics*, 2002, pp. 311–318.

[46] R. Vedantam, L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. European Conf. Computer Vision*, 2015, pp. 4566–4575.

[47] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. European Conf. Computer Vision*, 2016.

[48] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[49] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *Proc. Association Computational Linguistics*, vol. 2, 2014, pp. 67–78.

[50] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Proc. 52nd Annu. Meeting Association Computational Linguistics*, 2014, pp. 452–457.

[51] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Computer Vision*, 2015.

[52] Y. Cui, M. R. Ronchi, T.-Y. Lin, P. Dollár, L. Zitnick. (2015) COCO captioning challenge. [Online]. Available: http://mscoco.org/dataset/#captions-challenge

[53] Microsoft Cognitive Services Computer Vision API. [Online]. Available: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

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

[55] A. Agrawal, J. Lu, S. Antol, M. Mitchell, L. Zitnick, D. Batra, and D. Parikh, "VQA: Visual question answering," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 2425–2433.

[56] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[57] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. Int. Conf. Computer Vision*, 2017.

[58] T.-H. (K.). Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. Lawrence Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, "Visual storytelling," in *Proc. 2016 Conf. North American Chapter Association Computational Linguistics: Human Language Technologies*, 2016, pp. 1233–1239.

[59] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 30–42, Jan. 2012.

[60] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, and N. Jaitly, A, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, pp. 82–97, Dec. 2012.

[61] K. Koenigsbauer, Microsoft Office Blogs. (2016). [Online]. Available: https://blogs.office.com/2016/12/20/new-to-office-365-in-december-accessibility-updates-and-more/

[62] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese long short-term memory architecture for human re-identification," in *Proc. European Conf. Computer Vision*, 2016.

[63] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. European Conf. Computer Vision*, 2016.

[64] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *arXiv Preprint*, arXiv:1707.07998.

[65] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[66] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," *arXiv Preprint*, arXiv:1705.11001

[67] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[68] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. Association Advancement Artificial Intelligence*, 2017.

[69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* Cambridge, MA: MIT Press 2016.

[70] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and data sets," in *Computer Vision and Image Understanding*. Elsevier, 2017.

[71] Seeing AI. [Online]. Available: https://www.microsoft.com/en-us/seeing-ai/

[72] S. Reed, Z. Akata, X. Yan, L. Logeswaran, H. Lee, and B. Schiel "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Machine Learning*, 2016.

SP