Chapter 10
Deep Learning in Natural Language Generation from Images

Xiaodong He and Li Deng

Abstract Natural language generation from images, referred to as image or visual captioning also, is an emerging deep learning application that is in the intersection between computer vision and natural language processing. Image captioning also forms the technical foundation for many practical applications. The advances in deep learning technologies have created significant progress in this area in recent years. In this chapter, we review the key developments in image captioning and their impact in both research and industry deployment. Two major schemes developed for image captioning, both based on deep learning, are presented in detail. A number of examples of natural language descriptions of images produced by two state-of-the-art captioning systems are provided to illustrate the high quality of the systems' outputs. Finally, recent research on generating stylistic natural language from images is reviewed.

10.1 Introduction

In this final technical chapter of the book, we will discuss a very important but often lightly treated topic in natural language processing (NLP)—natural language generation (NLG), which had been progressing quite slowly until the recent rise of deep learning. As briefly discussed in Chap. 3 in the context of dialog systems, NLG is the process of generating text from a meaning representation and can be regarded as the reverse of natural language understanding.

In addition to serving as an integral component of dialog systems, NLG also plays a key role in text summarization, machine translation, image and video captioning, and other NLP applications. Both the earlier general-purpose rule-based and machine learning-based NLG systems were reviewed in Chap. 3, mainly for the specific dialog

X. He (✉)
Microsoft Research, Redmond, WA, USA
e-mail: xiaohe@microsoft.com

L. Deng
Citadel, Chicago & Seattle, USA
e-mail: l.deng@ieee.org

L. Deng and Y. Liu (eds.), *Deep Learning in Natural Language Processing*, https://doi.org/10.1007/978-981-10-5209-5_10

system application. In a few earlier chapters, more recent developments of deep learning-based methods for NLG, including mainly those based on recurrent neural nets and on the encoder–decoder deep neural architecture, were also briefly surveyed. These deep learning models can be trained from unaligned natural language data and can produce longer, more fluent utterances than previous methods.

In this chapter, rather than providing a comprehensive review of general NLG technology, we limit our scope to NLG in a special application—generating natural language sentences from images, or image captioning. This very difficult task had not been possible until deep learning methods for encoding images and for subsequent generation of natural language became matured within only past 2 years or so. The success of deep learning in image captioning presents another powerful evidence for the impact of deep learning in NLP in addition to several other NLP applications described in detail in the preceding chapters.

Generating a natural language description from an image or image captioning is an emerging interdisciplinary problem at the intersection of computer vision and NLP, and it 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, tremendous progress of this specialized NLG task has been achieved in recent years. In the remainder of this chapter, we will first summarize this exciting emerging NLG area, and then analyze the key development and the major progress. We will also discuss the impact of this progress both on research and on industry deployment, as well as potential future breakthroughs.

10.2 Background

It has been long envisioned that machines one day can understand the visual world at a human level of intelligence. Thanks to the progress in deep learning (Hinton et al. 2012; Dahl et al. 2011; Deng and Yu 2014), now researchers can build very deep convolutional neural networks (CNN), and achieve an impressively low error rate for tasks like large-scale image classification (Krizhevsky et al. 2012; He et al. 2015). In these tasks, to train a model for predicting the category of a given image, one can first annotate each image in a training set with a category 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 predominate object to be classified. The situation could be much more challenging when we want computers to understand complex scenes. Image captioning is one of such tasks. 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 relationships among them, and compose a coherent description about the overall content of the

image, which involves language and commonsense knowledge 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 with a lack of fine-grained alignments between the subregions in the image and the words in the description.

Further, unlike image classification tasks, where one 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 and to 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 proxies mainly for speeding up the development cycle of the system.

Early approaches to image captioning can be divided approximately into two families. The first one is based on template matching (Farhadi et al. 2010; Kulkarni et al. 2015). 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 (Hodosh et al. 2013; Ordonez et al. 2011). 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 could not 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 (Krizhevsky et al. 2012; He et al. 2015) and object detection (Girshick 2015) have motivated strong interest in using neural networks for visual captioning.

10.3 Deep Learning Frameworks to Generate Natural Language from an Image

10.3.1 The End-to-End Framework

Motivated by recent success of sequence-to-sequence learning in machine translation (Sutskever et al. 2014; Bahdanau et al. 2015), researchers studied an end-to-end encoder–decoder framework for image captioning (Vinyals et al. 2015; Karpathy and Fei-Fei 2015; Fang et al. 2015; Devlin et al. 2015; Chen and Zitnick 2015). Figure 10.1 illustrates a typical encoder–decoder-based captioning system (Vinyals

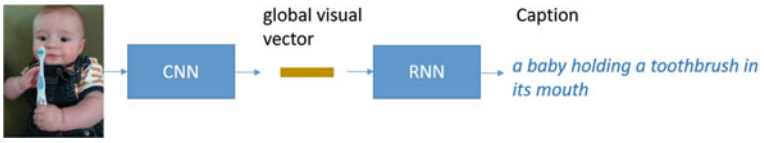

Fig. 10.1 NLG from an image using a CNN and RNN trained together in an end-to-end manner (figure from He and Deng 2017)

et al. 2015). In this framework, first the raw image is encoded by a global visual feature vector which represents the overall semantic information of the image, via deep CNN. As illustrated in Fig. 10.2, a CNN consists of several convolutional, max-pooling, response-normalization, and fully connected layers. Here, the CNN is trained for a 1000-class image classification task on the large-scale ImageNet dataset (Deng et al. 2009). The last layer of this AlexNet contains 1000 nodes, each corresponding to a category. Meanwhile, the second last fully connected dense layer is extracted as the global visual feature vector, representing the semantic content of the overall images. Given a raw image, the activation values at the second to the last fully connected layer are usually extracted as the global visual feature vector. This architecture has been very successful for large-scale image classification, and the learned features have shown to transfer to a broad variety of vision tasks.

Once the global visual vector is extracted, it is then fed into another recurrent neural network (RNN)-based decoder for caption generation, as illustrated in Fig. 10.3. 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. At the same time, 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 is generated. In practice, a long-short memory network (LSTM) (Hochreiter and Schmidhuber 1997) or gated recurrent unit (GRU) (Chung et al. 2015) variation of the RNN is often used, both of which have been shown to be more efficient and effective in training and capturing long-span language dependency (Bahdanau et al. 2015; Chung et al. 2015), and have found successful applications in action recognition tasks (Varior et al. 2016).

The representative set of studies using the above end-to-end framework include (Chen and Zitnick 2015; Devlin et al. 2015; Donahue et al. 2015; Gan et al. 2017a, b; Karpathy and Fei-Fei 2015; Mao et al. 2015; Vinyals et al. 2015) for image captioning and (Venugopalan et al. 2015a, b; Ballas et al. 2016; Pan et al. 2016; Yu et al. 2016) 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 Karpathy and Fei-Fei (2015), Mao et al. (2015), while the LSTM was used in (Vinyals et al. 2015). The visual feature vector was only fed into the RNN once at the first time step in Vinyals et al. (2015), while it was used at each time step of the RNN in Karpathy and Fei-Fei (2015). It is useful to point out that the

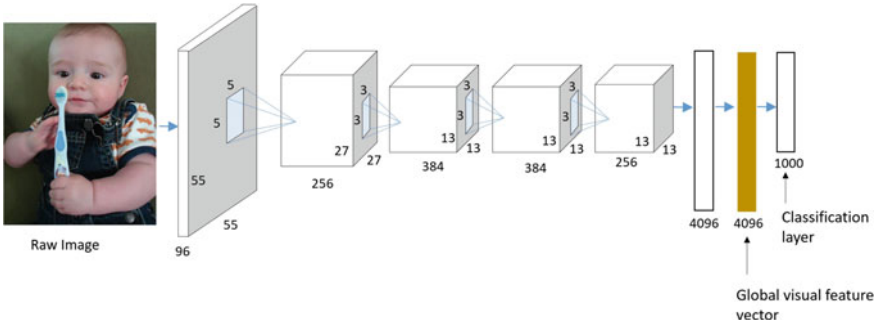

Fig. 10.2 A deep CNN (e.g., AlexNet) used as a front-end encoder of the image captioning system (figure from He and Deng 2017)

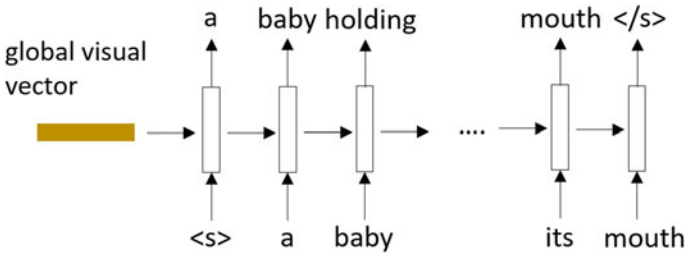

Fig. 10.3 An RNN used as a back-end decoder of the image captioning system (figure from He and Deng 2017)

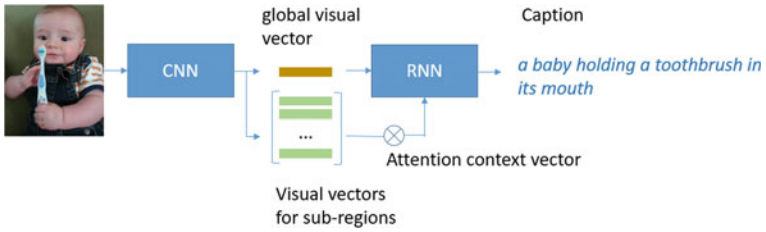

Fig. 10.4 The attention mechanism in the image captioning system's NLG process (figure from He and Deng 2017)

deep CNN, which is essential for the success of image-to-text applications described here, takes into account special translation-invariant properties of the image inputs.

Most recently, Xu et al. (2015) utilized an attention-based mechanism to learn where to focus in the image during caption generation. The attention architecture is illustrated in Fig. 10.4. 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 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 Yang et al. (2016), which introduced a review module to improve the attention mechanism and further by Liu et al. (2016), which proposed a method to improve the correctness of visual attention. More recently, based on object detection, a bottom-up attention model is proposed by Anderson et al. (2017), which demonstrates state-of-the-art performance on image captioning. In this framework, all the parameters, including the CNN, the RNN, and the attention model, can be trained jointly from the start to the end parts of the overall model; hence the name "end-to-end".

10.3.2 The compositional framework

Different from the end-to-end encoder–decoder framework just 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 10.5 illustrates a semantic-concept-detection-based compositional approach proposed by Fang et al. (2015). This approach is akin to and motivated by the long-standing architecture in speech recognition, consisting of multiple composed modules of the acoustic model, the pronunciation model, and the language model (Baker et al. 2009; Hinton et al. 2012; Deng et al. 2013; Deng and O'Shaughnessy 2003).

In this framework, the first step in the caption generation pipeline detects a set of semantic concepts, as known as tags or attributes, that are likely to be part of the images' 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

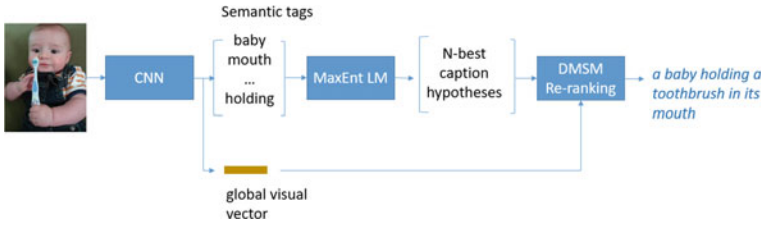

Fig. 10.5 A compositional approach based on semantic-concept-detection in image captioning (figure from He and Deng 2017)

the image bounding boxes corresponding to the words are unknown. To address this issue, Fang et al. (2015) proposed learning the detectors using the weakly supervised approach of multiple instance learning (MIL) (Zhang et al. 2005). While in Tran et al. (2016), this problem is treated as a multi-label classification task.

In Fang et al. (2015), 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 that defines the probability distribution over word sequences.

All these hypotheses were then re-ranked 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, a multimodal extension of the deep structured semantic model developed earlier for information retrieval (Huang et al. 2013). This "semantic" model 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 in system development and deployment, and facilitates exploiting various data sources to optimizing the performance of different modules more effectively, rather than learn all the models on limited image-caption paired data. On the other hand, end-to-end model usually has a simpler architecture and can optimize different components of the overall system jointly for a better performance.

More recently, a class of models have been proposed to integrate explicit semantic-concept-detection in an encoder–decoder framework. For example, Ballas et al. (2016) applied retrieved sentences as additional semantic information to guide the LSTM when generating captions, while Fang et al. (2015), You et al. (2016), Tran et al. (2016) applied a semantic-concept-detection process before generating sentences. In Gan et al. (2017b), a semantic compositional network is constructed based on the probability of detected semantic concepts for composing captions. This line of methods also represents the current state-of-the-art in image captioning.

From the architectural and task-definition points of view, this type of compositional framework for image captioning and for speech recognition shares a number of common themes. Both of the tasks have the output of natural language sentences, with different inputs of image pixels in the former and of speech waves in the latter. The attribute detection module in image captioning plays a similar role to the phonetic recognition module in speech recognition (Deng and Yu 2007). The use of language model to transform the detected attributes in the image to a list of caption hypotheses in image captioning has the correspondence in the later stages of speech recognition that turn the acoustic features and phonetic units into a collection of lexically correct word hypotheses (via a pronunciation model) and then into a linguistically plausible word sequence (via a language model) (Bridle et al. 1998; Deng 1998). The final, re-ranking module in image captioning is unique in that the earlier module of attribute detection does not possess the global information of the full image, while to

generate a meaningful natural sentence for the full image requires such information. In contrast, this requirement for matching global properties of input and output is not needed in speech recognition,

10.3.3 Other Frameworks

In addition to the two main frameworks for image captioning, other related frameworks also learn a joint embedding of visual features and associated captions, For example, Wei et al. (2015) have investigated to generate dense image captions for individual regions in images, and a variational autoencoder was developed in Pu et al. (2016) for image captioning. Further, motivated by the recent successes of reinforcement learning, image captioning researchers also proposed a set of reinforcement learning-based algorithms to directly optimize the captioning models for specific rewards. For example, Rennie et al. (2017) proposed a self-critical sequence training algorithm. It uses the REINFORCE algorithm to optimize an evaluation metric like CIDEr, which is usually not differentiable and therefore not easy to optimize by conventional gradient-based methods. In Ren et al. (2017), 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 the generative adversarial network (GAN) are proposed recently for text generation. Among them, SeqGAN (Yu et al. 2017) models the generator as a stochastic policy in reinforcement learning for discrete outputs like texts, and RankGAN (Lin et al. 2017) proposes 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.

10.4 Evaluation Metrics and Benchmarks

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 bilingual evaluation understudy BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam et al. 2015), and SPICE (Anderson et al. 2016). BLEU (Papineni et al. 2002) is widely used in machine translation and measures the fraction of N-grams (up to 4-gram) that are in common between a hypothesis and a reference or set of references. METEOR (Denkowski and Lavie 2014) instead measures unigram precision and recall, but extends exact word matches to include similar words based on WordNet synonyms and stemmed tokens. CIDEr (Vedantam et al. 2015) also measures the n-gram match between the caption hypothesis and the references, while the n-grams are weighted by TF-IDF. SPICE (Anderson et al. 2016), instead, measures the F1 score of semantic propositional content contained in image captions given the references, and therefore, it gives

the best correlation to human judgment. These automatic metrics can be computed efficiently, and therefore greatly speed up the development of image captioning algorithms. However, all of these automatic metrics are known to only roughly correlate with human judgment (Elliott and Keller 2014).

Researchers have created many datasets to facilitate the research of image captioning. The Flickr dataset (Young et al. 2014) and the PASCAL sentence dataset (Rashtchian et al. 2010) were created for facilitating the research of image captioning. More recently, Microsoft sponsored the creation of the COCO (Common Objects in Context) dataset (Lin et al. 2015), the largest image captioning dataset available to the public today. The availability of the large-scale datasets significantly prompted research in image captioning in the last several years. In 2015, about 15 groups participated in the COCO Captioning Challenge (Cui et al. 2015). The entries in the challenge are evaluated by human judgment. In the competition, all entries are assessed based on the results of M1—percentage of captions that are evaluated as better or equal to human caption, and M2—the percentage of captions that pass Turing test. Additional three metrics have been used as diagnostic and interpretation of the results: M3—Average correctness of the captions on a scale 1–5 (incorrect–correct), M4—average amount of detail of the captions on a scale 1–5 (lack of details—very detailed), and M5—percentage of captions that are similar to human description. More 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 are generated by human. If the judge chooses the automatically generated caption, or choose "cannot tell" option, it is considered to have passed Turing test.

The results, quantified by M1 to M5 metrics above, obtained from the top 15 image captioning systems in the 2015 COCO Captioning Challenge plus other recent top entries measured by automatic metrics have been summarized and analyzed in (He and Deng 2017). The success of these systems reflects the huge progress in this challenging task from perception to cognition achieved by deep learning methods.

10.5 Industrial Deployment of Image Captioning

Propelled by the fast progress in the research community, the industry started deploying image captioning services. In March 2016, Microsoft released the image captioning service as a cloud API to the public. To showcase the usage of the functionality, it also deployed a web application called CaptionBot (http://CaptionBot.ai), which captions arbitrary pictures users uploaded. More recently, Microsoft also deployed the caption service in the widely used product Office, specifically, Word and PowerPoint, for automatically generating alter-text for accessibility. Facebook also released an automatic image captioning tool that provides a list of objects and scenes identified in a photo. Meanwhile, Google open sourced their image captioning system

for the community (https://github.com/tensorflow/models/tree/master/im2txt), as a step toward public deployment of the captioning service.

With all these industrial-scale deployment and open-source projects, a massive number of images and user feedbacks in real-world scenarios are being collected that serve as the ever-increasing training data to steadfastly improve the performance of the systems. This will in turn stimulate new research in deep learning methods for visual understanding and natural language generation.

10.6 Examples: Natural Language Descriptions of Images

In this section, we provide typical examples of generating natural language captions that describe the contents of digital images, using the various deep learning techniques described in the preceding sections.

Given a digital image, such as a photo shown in the upper part of Fig. 10.6, the machine-generated textual description of the contents of the image—"a woman in a kitchen preparing food"—together with the human-annotated description—"woman working on counter near kitchen sink preparing a meal"—are shown in the lower part of the figure. In this case, an independent human (a mechanical Turker) slightly prefers the machine-generated text. Among the many images from Microsoft COCO database, about 30% of images are of this type, i.e., whose captions by the system are preferred, or are viewed equally good as human-generated captions.

From Figs. 10.7, 10.8, 10.9 and 10.10, we provide several other examples where mechanical Turkers prefer machine-generated textual descriptions of images to human-annotated ones, or view them as equally good.

The image captioning system that provides the above examples has been implemented in CaptionBot via calling Microsoft Cognitive Services, which allows mobile phone users to upload any photo from the phone to obtain its corresponding natural language caption. Several examples are provided from Figs. 10.11, 10.12 and 10.13. In the last example, we include the result when the celebrity detection component is added to the captioning system.

10.7 Recent Research on Generating Stylistic Natural Language from Images

The natural language captions generated by deep learning systems from images, with numerous techniques and examples provided in the preceding sections, usually gave only a factual description of the image content (Vinyals et al. 2015; Mao et al. 2015; Karpathy and Fei-Fei 2015; Chen and Lawrence Zitnick 2015; Fang et al. 2015; Donahue et al. 2015; Xu et al. 2015; Yang et al. 2016; You et al. 2016; Bengio et al. 2015; Tran et al. 2016). The natural language style has often been overlooked in the

Fig. 10.6 An example of image captioning in contrast with human annotation

Fig. 10.7 Another example of image captioning in contrast with human annotation

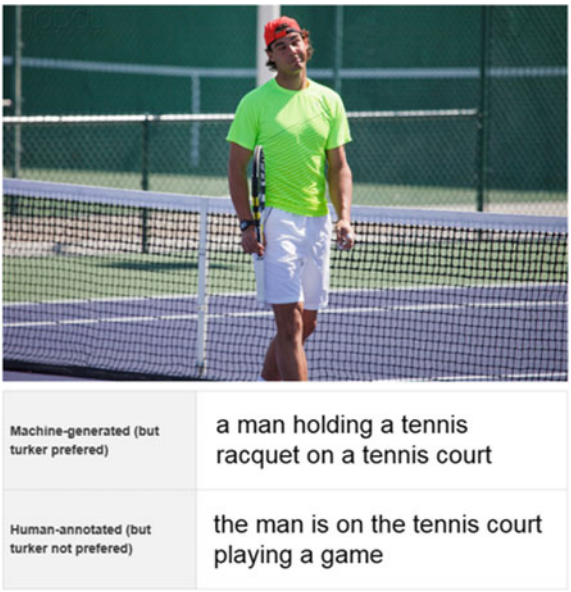

Fig. 10.8 Another example of image captioning in contrast with human annotation

caption generation process. Specifically, the existing image captioning systems have been using a language generation model that mixes the style with other linguistic patterns of language generation, thereby lacking a mechanism to control the style explicitly. The recent research aims to overcome this deficiency (Gan et al. 2017a) and is reviewed here.

A romantic or humorous natural language description of an image can greatly enrich the expressibility of the caption and make it more attractive. An attractive image caption will add more visual interest to images and can even become a distinguishing trademark of the captioning system. This is particularly valuable for certain applications; e.g., increasing user engagement in chatting bots or enlightening users in photo captioning for social media.

Gan et al. (2017a) proposed the StyleNet, which is able to produce attractive visual captions with styles only using monolingual stylized language corpus (i.e., without paired images) and standard factual image/video–caption pairs. StyleNet is built upon the recently developed methods that combine convolutional neural networks (CNNs) with recurrent neural networks (RNNs) for image captioning. The work is also motivated by the spirit of multitask sequence-to-sequence training Luong et al. (2015). Particularly, it introduces a novel factored LSTM model that can be used to disentangle the factual and style factors from the sentences through multitask training. Then at running time, the style factors can be explicitly incorporated to generate different stylized captions for an image.

The StyleNet has been evaluated on a newly collected Flickr stylized image caption dataset, with the results demonstrating that the proposed StyleNet significantly outperforms previous state-of-the-art image captioning approaches, measured by a

Machine-generated (but turker prefered)	a kitchen with wooden cabinets and a sink
Human-annotated (but turker not prefered)	an ornate kitchen is designed with rustic wooden parts

Fig. 10.9 Another example of image captioning in contrast with human annotation

Machine-generated (but turker prefered)	a clock tower in the middle of the street
Human-annotated (but turker not prefered)	a statue with a clock on it near a parking lot

Fig. 10.10 A final example of image captioning in contrast with human annotation

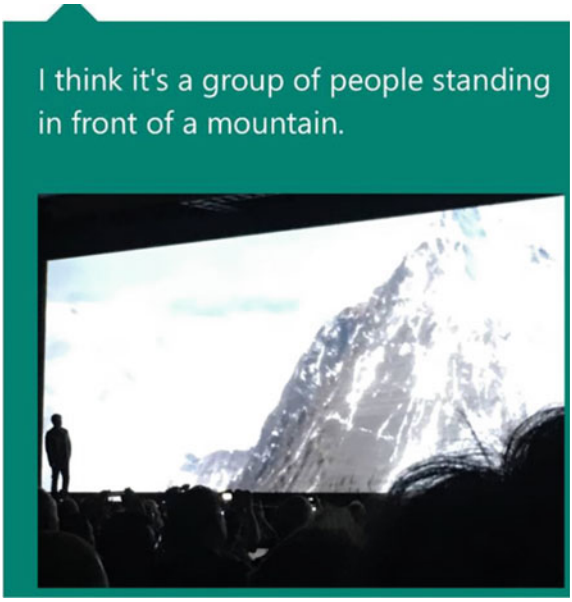

Fig. 10.11 The image which automatically generates natural sentence of "I think it's a group of people standing in front of a mountain." using Microsoft Cognition Services

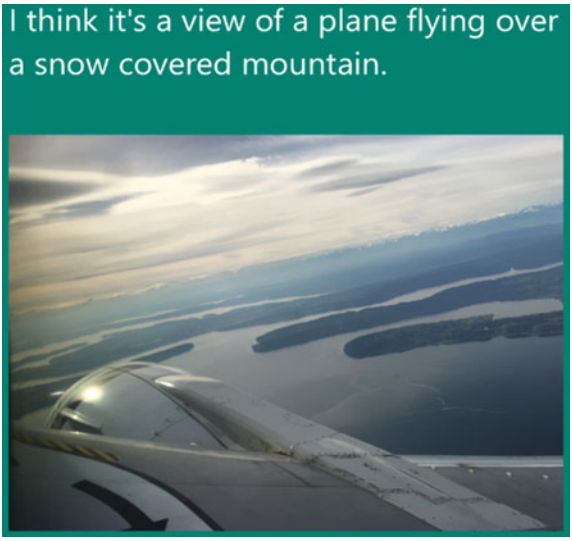

Fig. 10.12 The image which automatically generates natural sentence of "I think it's a view of a plane flying over a snow covered mountain." using Microsoft Cognition Services

"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background."

Fig. 10.13 The image which automatically generates a natural sentence using Microsoft Cognition Services with an added celebrity detection component

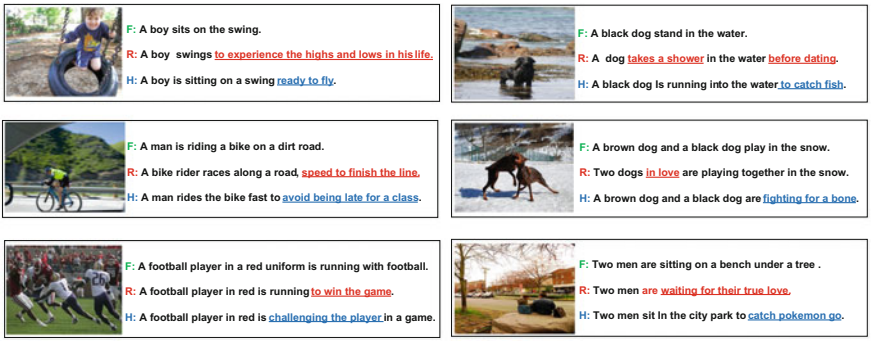

Fig. 10.14 Six examples of natural language captions generated by the StyleNet from images each with three different styles

set of automatic metrics and human evaluation. Some typical examples of stylistic caption generation are shown in Fig. 10.14, where it is observed that the caption with the standard factual style only describes the facts in the image in a dull language, while both the romantic and humorous style captions not only describe the content of the image but also express the content in a romantic or humorous way through generating phrases that bear a romantic (e.g., in love, true love, enjoying, dating, win the game, etc.) or humorous (e.g., find gold, ready to fly, catch Pokemon Go, bone, etc.) sense. Further, it has been found that the phrases that the StyleNet generates fit the visual content of the image coherently, making the caption visually relevant and attractive.

10.8 Summary

Natural language generation from images, or image captioning, is an emerging deep learning application that intersects computer vision and natural language processing. It also forms the technical foundation for many practical applications. Thanks to deep learning technologies, we have seen significant progress in this area in recent years. In this chapter, we have reviewed the key developments in image captioning that the community has made and their impact on both research and industry deployment. Two major frameworks developed for image captioning, both based on deep learning, are reviewed in detail. A number of examples of natural language descriptions of images produced by two state-of-the-art captioning systems are provided to illustrate the high quality of the systems' outputs.

Looking forward, while image captioning is a particular application of NLG in NLP, it is also a 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 (Fei-Fei and Perona 2016; Young et al. 2014; Agrawal et al. 2015), visual storytelling (Huang et al. 2016), visually grounded dialog (Das et al. 2017), and image synthesis from text description (Zhang et al. 2017). The progress in multimodal intelligence involving natural language is critical for building general artificial intelligence abilities in the future. The review provided in this chapter can hopefully encourage new students and researchers alike to contribute to this exciting area.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, L., Batra, D., & Parikh, D. (2015). Vqa: Visual question answering. In *ICCV*.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *ECCV*.

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

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Baker, J., et al. (2009). Research developments and directions in speech recognition and understanding. *IEEE Signal Processing Magazine*, *26*(4),

Ballas, N., Yao, L., Pal, C., & Courville, A. (2016). Delving deeper into convolutional networks for learning video representations. In *ICLR*.

Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS* (pp. 1171–1179).

Bridle, J., et al. (1998). An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. *Final Report for 1998 Workshop on Language Engineering, Johns Hopkins University CLSP*.

Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *CVPR* (pp. 2422–2431).

Chen, X., & Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. In *ICML*.
Cui, Y., Ronchi, M. R., Lin, T. -Y., Dollar, P., & Zitnick, L. (2015). Coco captioning challenge. In http://mscoco.org/dataset/captions-challenge2015.
Dahl, G., Yu, D., & Deng, L. (2011). Large-vocabulry continuous speech recognition with context-dependent DBN-HMMs. In *Proceedings of ICASSP*.
Das, A., et al. (2017). Visual dialog. In *CVPR*.
Deng, L. (1998). A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, *24*(4),
Deng, L., & O'Shaughnessy, D. (2003). *SPEECH PROCESSING A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker.
Deng, L., & Yu, D. (2007). Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition. In *Proceedings of ICASSP*.
Deng, L., & Yu, D. (2014). *Deep Learning: Methods and Applications*. Breda: NOW Publishers.
Deng, J., Dong, W., Socher, R., Li, L. -J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255).
Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *Proceedings of ICASSP*.
Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *ACL*.
Devlin, J., et al. (2015). Language models for image captioning: The quirks and what works. In *Proceedings of CVPR*.
Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *CVPR* (pp, 2625–2634).
Elliott, D., & Keller, F. (2014). Comparing automatic evaluation measures for image description. In *ACL*.
Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., et al. (2015). From captions to visual concepts and back. In *CVPR* (pp. 1473–1482).
Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *ECCV*.
Fei-Fei, L., & Perona, P. (2016). Stacked attention networks for image question answering. In *Proceedings of CVPR*.
Gan, C., et al. (2017a). Stylenet: Generating attractive visual captions with styles. In *CVPR*.
Gan, Z., et al. (2017b). Semantic compositional networks for visual captioning. In *CVPR*.
Girshick, R. (2015). Fast r-cnn. In *ICCV*.
He, X., & Deng, L. (2017). Deep learning for image-to-text generation. In *IEEE Signal Processing Magazine*.
He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In *CVPR*.
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. -r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., & Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*.
Huang, P., et al. (2013). Learning deep structured semantic models for web search using clickthrough data. *Proceedings of CIKM*.
Huang, T. -H., et al. (2016). Visual storytelling. In *NAACL*.
Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR* (pp. 3128–3137).

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*.
Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2015). Babytalk: Understanding and generating simple image descriptions. In *CVPR*.
Lin, K., Li, D., He, X., Zhang, Z., & Sun, M.- T. (2017). Adversarial ranking for language generation. In *NIPS*.
Lin, T. -Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollar, P. (2015). Microsoft coco: Common objects in context. In *ECCV*.
Liu, C., Mao, J., Sha, M., & Yuille, A. (2016). Attention correctness in neural image captioning. preprint arXiv:1605.09553.
Luong, M. -T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. In *ICLR*.
Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2015). Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*.
Ordonez, V., Kulkarni, G., Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *NIPS*.
Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. In *CVPR*.
Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *ACL* (pp. 311–318).
Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In *NIPS*.
Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using amazons mechanical turk. In *NAACL HLT Workshop Creating Speech and Language Data with Amazons Mechanical Turk*.
Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. -J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*.
Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *CVPR*.
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS* (pp. 3104–3112).
Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., & Sienkiewicz, C. (2016). Rich image captioning in the wild. arXiv preprint arXiv:1603.09016.
Varior, R. R., Shuai, B., Lu, J., Xu, D., & Wang, G. (2016). A siamese long short-term memory architecture for human re-identification. In *ECCV*.
Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *CVPR* (pp. 4566–4575).
Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015a). Sequence to sequence-video to text. In *ICCV*.
Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2015b). Translating videos to natural language using deep recurrent neural networks. In *NAACL*.
Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR* (pp. 3156–3164).
Wei, L., Huang, Q., Ceylan, D., Vouga, E., & Li, H. (2015). Densecap: Fully convolutional localization networks for dense captioning. *Computer Science*.
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML* (pp. 2048–2057).
Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., & Cohen, W. W. (2016). Encode, review, and decode: Reviewer module for caption generation. In *NIPS*.
You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *CVPR*.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of ACL*.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*.

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*.

Zhang, H., et al. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*.

Zhang, C., Platt, J. C., & Viola, P. A. (2005). Multiple instance boosting for object detection. In *NIPS*.