Neurocomputing 311 (2018) 291–304

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A survey on automatic image caption generation

Shuang Bai [a,*], Shan An [b]

[a] School of Electronic and Information Engineering, Beijing Jiaotong University, No.3 Shang Yuan Cun, Hai Dian District, Beijing, China
[b] Beijing Jingdong Shangke Information Technology Co., Ltd, Beijing, China

ARTICLE INFO

Article history:
Received 5 May 2017
Revised 13 April 2018
Accepted 19 May 2018
Available online 26 May 2018

Communicated by Dr. Min Xu

Keywords:
Image captioning
Sentence template
Deep neural networks
Multimodal embedding
Encoder–decoder framework
Attention mechanism

ABSTRACT

Image captioning means automatically generating a caption for an image. As a recently emerged research area, it is attracting more and more attention. To achieve the goal of image captioning, semantic information of images needs to be captured and expressed in natural languages. Connecting both research communities of computer vision and natural language processing, image captioning is a quite challenging task. Various approaches have been proposed to solve this problem. In this paper, we present a survey on advances in image captioning research. Based on the technique adopted, we classify image captioning approaches into different categories. Representative methods in each category are summarized, and their strengths and limitations are talked about. In this paper, we first discuss methods used in early work which are mainly retrieval and template based. Then, we focus our main attention on neural network based methods, which give state of the art results. Neural network based methods are further divided into subcategories based on the specific framework they use. Each subcategory of neural network based methods are discussed in detail. After that, state of the art methods are compared on benchmark datasets. Following that, discussions on future research directions are presented.

1. Introduction

Humans are able to relatively easily describe the environments they are in. Given an image, it is natural for a human to describe an immense amount of details about this image with a quick glance [1]. This is one of humans' basic abilities. Making computers imitate humans' ability to interpret the visual world has been a long standing goal of researchers in the field of artificial intelligence.

Although great progress has been made in various computer vision tasks, such as object recognition [2], [3], attribute classification [4], [5], action classification [6], [7], image classification [8] and scene recognition [9], [10], it is a relatively new task to let a computer use a human-like sentence to automatically describe an image that is forwarded to it.

Using a computer to automatically generate a natural language description for an image, which is defined as image captioning, is challenging. Because connecting both research communities of computer vision and natural language processing, image captioning not only requires a high level understanding of the semantic contents of an image, but also needs to express the information in a human-like sentence. Determination of presences, attributes and relationships of objects in an image is not an easy task itself. Organizing a sentence to describe such information makes this task even more difficult.

Since much of human communication depends on natural languages, whether written or spoken, enabling computers to describe the visual world will lead to a great number of possible applications, such as producing natural human robot interactions, early childhood education, information retrieval, and visually impaired assistance, and so on.

As a challenging and meaningful research field in artificial intelligence, image captioning is attracting more and more attention and is becoming increasingly important.

Given an image, the goal of image captioning is to generate a sentence that is linguistically plausible and semantically truthful to the content of this image. So there are two basic questions involved in image captioning, i.e. visual understanding and linguistic processing. To ensure generated sentences are grammatically and semantically correct, techniques of computer vision and natural language processing are supposed to be adopted to deal with problems arising from the corresponding modality and integrated appropriately. To this end, various approaches have been proposed.

Originally, automatic image captioning is only attempted to yield simple descriptions for images taken under extremely constrained conditions. For example, Kojima et al. [11] used concept hierarchies of actions, case structures and verb patterns to generate natural languages to describe human activities in a fixed office

* Corresponding author.
E-mail address: shuangb@bjtu.edu.cn (S. Bai).

https://doi.org/10.1016/j.neucom.2018.05.080
0925-2312/

Table 1
Summary of image captioning methods.

Method		Representative methods
Early work	Retrieval based	Farhadi et al. [13], Ordonez et al. [15], Gupta et al. [16], Hodosh et al. [32], Mason and Charniak [49], Kuznetsova et al. [50].
	Template based	Yang et al. [14], Kulkarni et al. [51], Li et al. [52], Mitchell et al. [53], Ushiku et al. [54].
Neural networks based	Augmenting early work by deep models	Socher et al. [55], Karpathy et al. [37], Ma et al. [56], Yan and Mikolajczyk [57], Lebret et al. [58].
	Multimodal learning	Kiros et al. [59], Mao et al. [60], Karpathy and Li [61], Chen and Zitnick [62].
	Encoder–decoder framework	Kiros et al. [63], Vinyals et al. [64], Donahue et al. [34], Jia et al. [65] Wu et al. [66], Pu et al. [67].
	Attention guided	Xu et al. [68], You et al. [69], Yang et al. [70].
	Compositional architectures	Fang et al. [33], Tran et al. [71], Fu et al. [72], Ma and Han [73], Oruganti et al. [74], Wang et al. [75].
	Describing novel objects	Mao et al. [76], Hendricks and Venugopalan, [36].

environment. Hede et al. used a dictionary of objects and language templates to describe images of objects in backgrounds without clutters [12]. Apparently, such methods are far from applications to describing images that we encounter in our everyday life.

It is not until recently that work aiming to generate descriptions for generic real life images is proposed [13]–[16]. Early work on image captioning mainly follows two lines of research, i.e. retrieval based and template based. Because these methods accomplish the image captioning task either by making use of existing captions in the training set or relying on hard-coded language structures, the disadvantage of methods adopted in early work is that they are not flexible enough. As a result, expressiveness of generated descriptions by these methods is, to a large extent, limited.

Despite the difficult nature of the image captioning task, thanks to recent advances in deep neural networks [17–22], which are widely applied to the fields of computer vision [23–26] and natural language processing [27–31], image captioning systems based on deep neural networks are proposed. Powerful deep neural networks provide efficient solutions to visual and language modelling. Consequently, they are used to augment existing systems and design countless new approaches. Employing deep neural networks to tackle the image captioning problem has demonstrated state of the art results [32]–[37].

With the recent surge of research interest in image captioning, a large number of approaches have been proposed. To facilitate readers to have a quick overview of the advances of image captioning, we present this survey to review past work and envision future research directions. Although there exist several research topics that also involve both computer vision and natural language processing, such as visual question answering [38–42], text summarization [43], [44] and video description [45–48], because each of them has its own focus, in this survey we mainly focus on work that aims to automatically generate descriptions for generic real life images.

Based on the technique adopted in each method, we classify image captioning approaches into different categories, which are summarized in Table 1. Representative methods in each category are listed. Methods in early work are mainly retrieval and template based, in which hard coded rules and hand engineered features are utilized. Outputs of such methods have obvious limitations. We review early work relatively briefly in this survey. With the great progress made in research of deep neural networks, approaches that employ neural networks for image captioning are proposed and demonstrate state of the art results. Based on the framework used in each deep neural network based method, we further classify these methods into subcategories. In this survey, we will focus our main attention on neural network based methods. The framework used in each subcategory will be introduced, and the corresponding representative methods will be discussed in more detail.

This paper is organized as follows. In Sections 2 and 3, we first review retrieval based and template based image captioning methods, respectively. Section 4 is about neural network based methods, in this section we divide neural network based image captioning methods into subcategories, and discuss representative methods in each subcategory, respectively. State of art methods will be compared on benchmark datasets in Section 5. After that, we will envision future research directions of image captioning in Section 6. The conclusion will be given in Section 7.

2. Retrieval based image captioning

One type of image captioning methods that are common in early work is retrieval based. Given a query image, retrieval based methods produce a caption for it through retrieving one or a set of sentences from a pre-specified sentence pool. The generated caption can either be a sentence that has already existed or a sentence composed from the retrieved ones. First, let us investigate the line of research that directly uses retrieved sentences as captions of images.

Farhadi et al. establish a ⟨*object, action, scene*⟩ meaning space to link images and sentences. Given a query image, they map it into the meaning space by solving a Markov Random Field, and use Lin similarity measure [77] to determine the semantic distance between this image and each existing sentence parsed by Curran et al. parser [78]. The sentence closest to the query image is taken as its caption [13].

In [15], to caption an image Ordonez et al. first employ global image descriptors to retrieve a set of images from a web-scale collection of captioned photographs. Then, they utilize semantic contents of the retrieved images to perform re-ranking and use the caption of the top image as the description of the query.

Hodosh et al. frame image captioning as a ranking task [32]. The authors employ the Kernel Canonical Correlation Analysis technique [79], [80] to project image and text items into a common space, where training images and their corresponding captions are maximally correlated. In the new common space, cosine similarities between images and sentences are calculated to select top ranked sentences to act as descriptions of query images.

To alleviate impacts of noisy visual estimation in methods that depend on image retrieval for image captioning, Mason and Charniak first use visual similarity to retrieve a set of captioned images for a query image [49]. Then, from the captions of the retrieved images, they estimate a word probability density conditioned on the query image. The word probability density is used to score the existing captions to select the one with the largest score as the caption of the query.

The above methods have implicitly assumed that given a query image there always exists a sentence that is pertinent to it. This assumption is hardly true in practice. Therefore, instead of using retrieved sentences as descriptions of query images directly, in the other line of retrieval based research, retrieved sentences are utilized to compose a new description for a query image.