Automation of caption generation with AI

Ashwin Sadanandan Nambiar, Hemant Gautham Bodhare
Guide:Prof Divya Premchandran
Keraleeya Smajam's Model College, khambalpada Road Thakurli,Dombivili (East)

Abstract—– Automatically creating a caption for a picture is known as image captioning. Itis becoming more popular as a freshly developed scientific field. The semantic content of images must be gathered and communicated in natural languages in order to accomplish the purpose of image captioning. Image captioning is a difficult task since it bridges the computer vision and natural language processing research fields. Different strategies have been put forth to address this issue. We give a survey of developments in picture captioning research in this publication. We categorize various techniques to image captioning into distinct groups based on the strategy used. Each category's representative approaches are outlined along with their advantages and disadvantages. In this paper, we first go over retrieval and template-based strategies that were often employed in earlier research. Then, as these techniques produce state-of-the-art outcomes, we concentrate mostly on neural network-based approaches. Based on the particular framework theyemploy, neural network-based solutions are further subdivided. There is a detailed discussion of each subcategory of neural network-based techniques. Following that, benchmark datasets are used to compare state-of-the-art approaches. The discussion of potential directions for further study isthen offered.

Keywords — Image captioning, Deep neural network, VGGNet, RNN-LSTM, CNN

1.INTRODUCTION

Humans are able to characterize their surroundings quite quickly. When presented with an image, it is normal for a human to quickly explain a vast quantity of details about the image [6]. Oneof the fundamental human capacities is this. Researchers in the field of artificial intelligence have long sought to make computers mimic humans' understanding of the visual environment. Despite significant advancements in a number of computer vision tasks, including object recognition [7], [9], attribute classification [15], [8], action classification [17], [3], image classification [14], and scene recognition [23], [8], it is still a relatively new task to allow a computer to automatically describe an image that has been sent to it using a sentence that sounds human. Image captioning is the processof automatically creating a natural language description for an image using a computer. Because it bridges the computer vision and natural language processing research areas, image captioning calls for not only a deep comprehension of the semantic contents of an image but also the ability to articulate the information in human-like sentences. Finding the presence, characteristics, and relationships of objects in an image is a difficult task in and of itself. This endeavor is made more challenging by the organization required to describe such material in a sentence. Giving computers the ability to describe the visual world will open up a wide range of potential applications, including the creation of natural human-robot interactions, early childhood education, information retrieval, and assistance for those who are visually impaired, among many others. This is because a large portion of human communication, whether written or spoken, depends on natural languages. Image captioning is gaining popularity as a difficult and significant area of artificial intelligence study and is becoming more and more crucial. The objective of picture captioning is to produce a statement that, given an image, is both linguistically and semantically true to the image's content. Therefore, there are two fundamental issues with image captioning: language processing and visual perception. Techniques of computer vision and natural language processing should be implemented to address issues originating from the associated modality and integrated effectively to guaranteethat generated sentences are grammatically and semantically correct. Various strategies have been put forth to this purpose. Automatic image captioning was initially just intended to produce briefcaptions for photos shot under relatively limited circumstances. For instance, Kojima et al. [13] generated natural languages to express human activities in a fixed office environment using idea hierarchies of actions, case structures, and verb patterns. To characterize photographs of things on backgrounds free of clutter, Hede et al. employed a vocabulary of objects and language templates[11]. Evidently, these techniques are not appropriate for describing visuals that we see every day. Work to produce descriptions for generic real-world photographs has just recently been proposed. The early work on captioning for images generally adheres to two lines of inquiry: retrieval-basedand template-based. The drawback of approaches used in early work is that they are not flexible enough because they either rely on hard-coded language structures or use existing captions fromthe training set to complete the process of image captioning. As a result, the expressiveness of descriptions generated by these methods is largely constrained. The task of captioning images is challenging, but recent developments in deep neural networks, which are widely used in computervision and natural language processing, have made it possible to propose deep neural network-based image captioning systems. Effective solutions for visual and linguistic modelling are provided by powerful deep neural networks. As a result, they are employed to improve current systems and create a vast array of fresh ideas. Deep neural networks have produced state-of-the-art results when used to solve the picture captioning problem . Many different strategies have been put forth in response to the recent increase in scholarly interest in image captioning. We present this survey to allow readers to quickly examine the developments in picture captioning while also considering potential future research topics. Although there are many research areas that combine computer vision and NLP, including visual question answering, text summarization , and video description, each of them has a specific focus, so in this survey we primarily concentrate on work that aims to automatically generate descriptions for common real-world images. We categorize picture captioning methods into various groups according to the strategy used in eachone. These groups are listed in Table 1. There is a list of illustrative techniques for each category.Early research primarily used retrieval and template-based methods, where hard-coded rules and manually engineered features were used. These methodologies' results clearly have some limits. In this survey, we fairly quickly review early work. With the significant advancements achieved in the study of deep neural networks, methods that use neural networks for image captioning aredeveloped and show cutting-edge outcomes. We further divide these methods into subcategories based on the framework each deep neural network-based method uses. We will pay particular emphasis to neural network-based techniques in this survey. Each subcategory's framework will be introduced, and further discussion of the relevant representative methods, will follow. We initially discuss retrieval-based and template-based picture captioning techniques in Sections 2 and 3, respectively. Neural network-based approaches are the subject of Part 4. In this section, we categorize neural network-based image captioning techniques into subcategories and describe exemplary techniques in each category individually. In Section 6, we will consider potential future prospects for picture captioning research. Section 7 will provide the conclusion.

1.Content

1.1IMAGE CAPTIONING BASED ON RETRIEVAL

Retrieval-based methods for captioning images were popular in earlier research. Retrieval-based approaches, given a query image, generate a caption for it by selecting one or more sentences froma pre-defined sentence pool. The generated caption may consist of a previously written statementor one that was put together using the recovered sentences. Let's look into the study avenue that directly uses recovered texts as image descriptions first. In order to connect language and images,Farhadi et al. create an object, action, and scene meaning space. They employ Lin similarity measure to calculate the semantic distance between a query image and each existing sentence processed by Curran et al. parser after mapping the query image into the meaning space using a Markov Random Field and Lin similarity measure. The caption for the image in question is determined by which phrase is closest to it [5]. In order to caption an image in [19], Ordonez et al. first obtain a series of images from a web-scale database of captioned photographs using global image descriptors. The caption of the top image is used as the description of the query, andthey then execute re-ranking using the semantic contents of the photos that were retrieved. Image captioning is framed as a rating assignment by Hodosh et al. [12]. In order to project picture and text elements into a shared space where training images and their corresponding captions are maximally connected, the authors use the Kernel Canonical Correlation Analysis approach [1]. The top-ranked sentences are chosen in the new common area to serve as descriptions of thequery photos based on the cosine similarity between the images and the sentences. Mason and Charniak initially employ visual similarity to retrieve a set of captioned images for a query imagein order to mitigate the effects of noisy visual estimate in systems that rely on image retrieval forimage captioning [18]. They then estimate a word probability density conditioned on the query image from the captions of the retrieved images. The current captions are scored using the word probability density, and the one with the highest score is chosen to serve as the query's caption. The aforementioned techniques have made the implicit assumption that, given a query image, a relevant text will always exist. In reality, this presumption hardly ever holds true. As a result, in the other branch of retrieval-based research, retrieved phrases are used to create a new descriptionfor a query image rather than using them directly as descriptions of query images. Gupta et al. process the sentences in the dataset using the Stanford Core NLP toolkit1 after being given a dataset of sentences and paired images. The result is a list of keywords for each image. Image retrieval is first carried out based on global image features to get a group of images for the query inorder to build a description for an image. The keywords associated with the retrieved photos are then chosen using a model trained to predict phrase relevance. The created description sentence is then based on the chosen pertinent phrases [10]. Similar in concept, Kuznetsova et al presented a tree-based method for creating visual descriptions from web photos with captions. Following phrase extraction from the images and model description composition, the authors encode the constraint optimization issue using integer linear programming and solve it using the CPLEX solver2. The same authors presented a comparable strategy in , which was published prior to thispaper. There are clear drawbacks to retrieval-based image captioning techniques. These techniques transmit properly constructed human-written sentences or phrases to create descriptions for image queries. Although the resulting outputs are frequently grammatically sound and fluid, limiting visual descriptions to preexisting words prevents them from adapting to unique object combinations or inventive scenarios. In other circumstances, created descriptions might even be irrelevant to the contents of the images. The ability of retrieval-based approaches to characterize images is severelyconstrained.

1.2Image Captioning Using Template

Another approach that is frequently employed in early image captioning efforts is template-based. Image captions are created using template-based methods using a procedure that is syntactically and semantically limited. Typically, a predetermined set of visual concepts must be detectedin order to employ a template-based method to construct a description for an image. Then, sentences are created using phrase templates, specialized language grammar rules, or combinatorial optimization algorithms , [21] which connect the visual notions that have been discovered. Yang etal. offer a technique in which a quadruplet (Nouns-Verbs-Scenes-Prepositions) is used as a sentence template for producing image descriptions in [22]. The authors estimate the items and scenes in this image using detection methods [7], before providing an image description. After that, they utilize a language mode trained on the Gigaword corpus3 to predict the possible verbs, situations,and prepositions to be included in the phrase. Using Hidden Markov Model inference, the optimal quadruplet is discovered after computing the probability of each element. In order to construct the visual description, the quadruplet's phrase structure is filled. Conditional Random Field is used by Kulkarni et al. to determine the image contents that will be displayed in the image caption [4].According to their method, a graph's nodes represent various types of objects, object properties, and spatial interactions between objects. In the graph model, pairwise potential functions are derived by doing statistics on a collection of existing descriptions, whereas unary potential functions of nodes are obtained by employing corresponding visual models. Conditional Random Field inference is used to determine the description of an image's contents. A description is generated using the inference's outputs and a sentence template. Visual models are used by Li et al. to do picture detections for obtaining objects, attributes, and other semantic information relationships in space . Then, for encoding recognition results, they define a triplet of the following format: adj1, obj1, prep, adj2, obj2. Using web-scale n-grams, create a description with the triplet. Data that can calculate frequency counts for potential n-grams sequences are used to carry out phrase selection such that it is possible to compile potential phrases for the triplet. Following that, phrasefusion is used to use dynamic programming to determine the ideal compatible group of phrases to serve as the description of the questioned picture. In order to interpret an image and represent it, Mitchell et al. use triplets of objects, actions, and spatial relationships. Then, based on the outcomes of the visual recognition, they create an image description as a tree-generating process. The writers select the description of an image's contents by grouping and ranking object nouns. Inorder to generate entire trees, object nouns are employed to create sub-trees, which are then used.Finally, a string from the generated full trees is chosen as the description of the matching image using a trigram language model. With the methods described above, words from a query image are piecemeal predicted using visual models. In later rounds, expected words like prepositions, objects, qualities, and verbs are joined to create descriptions that are human-like. Phrases contain larger bits of information than individual words since they are word combinations. Phrases that producesentences usually have a stronger descriptive element. So, approaches based on the framework forpicture captioning that uses templates are suggested. In order to directly develop phrase classifiers for image captioning, Ushiku et al. introduce a technique called Common Subspace for Model and Similarity [54]. The authors specifically extract continuous words [84] as phrases from training captions. Then, they combine similarity-based and model-based classification to learn a classifierfor each phrase by mapping picture features and phrase features into the same subspace. In the testing phase, multi-stack beam search is used to connect terms estimated from a query imagein order to build a description. The descriptions produced by template-based picture captioning can provide syntactically valid phrases, and they are typically more pertinent to the contents of the images than those produced by retrieval-based methods. Template-based techniques can havecertain drawbacks, too. There are frequently restrictions on coverage, creativity, and complexity of generated sentences because description creation under the template-based architecture is firmly confined to image contents recognized by visual models, given the typically low number of visualmodels available. Additionally, compared to captions written by humans, automated descriptions would seem less natural if sentences are constructed using inflexible templates.

1.3Captioning of Image using Deep Neural Network

Early work mostly adopts retrieval-based and template-based image captioning techniques. Recent work starts to rely on deep neural networks for automatic picture captioning because deep learning has advanced significantly [2], [16]. These techniques will be examined in this section. Even though deep neural networks are now frequently used to handle the task of captioning images, various approaches might be based on various frameworks. As a result, we divide deep neural network-based methods into subcategories based on the primary frameworks they employ and then talk about each group individually.

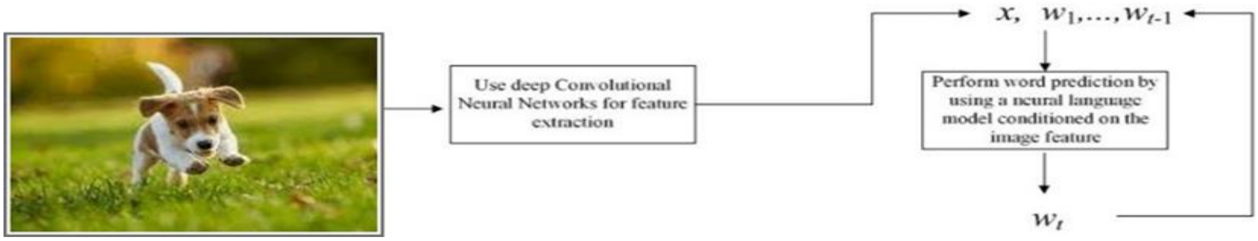

Use deep Convolutional Neural Networks for feature extraction
$x,\ w_1,\dots,w_{t-1}$
Perform word prediction by using a neural language model conditioned on the image feature
w_t

Figure 1: General Outline For Image Captioning

1.4Neural network-enhanced retrieval and template-based techniques

As a result of advancements in the field of deep neural networks, deep neural networks are now used to do picture captioning rather than hand-engineered features and shallow models as in earlier studies. Researchers propose to use deep models to structure image captioning as a multi-modality, drawing inspiration from retrieval-based approaches. Ranking issue and embedding According to a proposal made by Socher et al., dependency-tree recursive neural networks might be used to encode phrases and sentences as compositional vectors in order to extract a descriptive sentence for a query image. To extract characteristics from images [20], they employ a different deep neural network as a visual model. A max-margin objective function is used to map the obtained multimodal data into a common space. After training, the inner products of the right image and sentence pairings in the common area will be larger, and the opposite will be true. Finally, sentence retrieval is carried out using similarities found in the representations of sentences and images in the common space. For the purpose of ranking sentences for a query image, Karpathy et al. suggest to integrate sentence fragments and image fragments into a single space . Theyuse detection results from the Region Convolutional Neural Network approach [3] in an imageas image fragments and dependency tree relations [93] in a sentence as sentence fragments. The authors create a structured max margin objective, which consists of a global ranking term anda fragment alignment term, to map visual and literary data into a shared space by representing both phrase and image fragments as feature vectors. Sentence rating can be done at a finer level because in the common space, similarities between sentences and images are calculated based on fragment similarities. Ma et al. suggest a multimodal Convolutional Neural Network [56] to detect similarities between phrases and images while taking into account various levels of interaction between them. The three types of components in Ma's framework are multilayer perceptions to rate the compatibility of visual and textual data, matching CNNs to jointly represent visual and textual data, and image CNNs to encode visual data , respectively. To accommodate for joint representations of images and words, phrases, and sentences, the authors employ several matching CNN versions. An ensemble of multimodal Convolutional Neural Networks is used to determine the final matching score between a picture and a phrase. To match images and phrases [57], Yan and Mikolajczyk suggest using deep Canonical Correlation Analysis . They employ a stacked network to extract textual information from sentences with Frequency-Inverse Document Frequency representation, and a deep convolutional neural network [8] to extract visual data from photos. With the correlation between paired characteristics maximized, the canonical correlation analysis goal is used to map visual and textual information to a shared latent space. Similarities between an image feature and a sentence feature can be computed directly for sentence retrieval in the joint latent space. The use of deep models within the template-based framework is also attempted,in addition to using deep models to supplement retrieval-based picture captioning techniques. Lebret et al. use deep models to produce image descriptions using a form of soft-template. In this approach, the authors take phrases from training sentences using the SENNA software4 andcreate statistics using the words they have taken. Using word vector representation techniques, phrases are represented as high-dimensional vectors , while images are represented using a deep convolutional neural network . Given a query image, phrases can be inferred from it using a bilinear model that has been trained as a metric between image features and phrase features. Under the supervision of early stage statistics, phrases inferred from an image are employed to produce a sentence. Deep neural networks have a huge impact on how well image captioning algorithms perform. Deep neural networks are included into retrieval-based and template-based approaches, however this does not eliminate their drawbacks. These techniques do not remove the constraintson the phrases they produce.

1.1.1Multi modal Learning-Based Image Captioning

Generated sentences are constrained by image captioning techniques that rely on retrieval and templates. Powerful deep neural networks have made it possible to propose picture captioning methods that do not rely on preexisting captions or sentence structure presumptions throughout the caption generation process. Such techniques can produce sentences with deeper structures that are more expressive and adaptable. One method for creating image captions that relies solelyon learning is the use of multimodal neural networks. Fig. 1 depicts the general organization of multimodal learning-based picture captioning techniques. Such methods start by extracting picture features using a feature extractor, like deep convolutional neural networks. The resulting picturefeature is then passed on to a neural language model, which maps it into the shared space with word features and performs word prediction based on the image feature and previously created context words. In order to create captions for photographs, Kiros et al. suggest using a neural language model that is trained on image inputs. Their approach adapts the log-bilinear languagemodel to multimodal circumstances. A language model is used to predict the likelihood of creatinga word wt based on previously created words w1,..., wt-1 in an issue involving natural language processing, as shown below:

$$P\,(W_t|W_1.........W_{t-1}) \qquad (1)$$

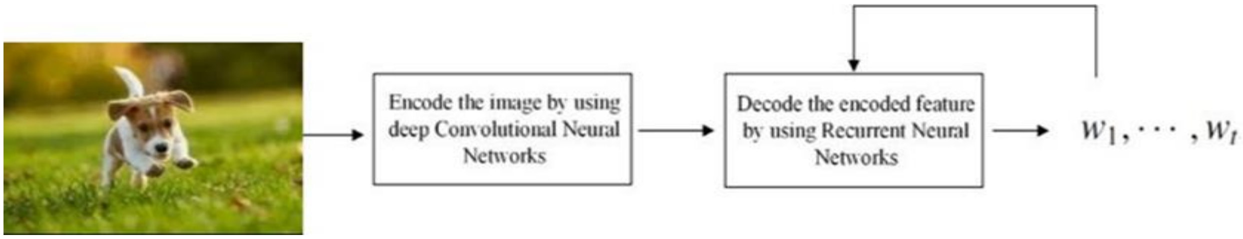

Encode the image by using deep Convolutional Neural Networks
Decode the encoded feature by using Recurrent Neural Networks
$W_1, \cdots, W_t$

Figure 2: The fundamental Design For Encoder Decoder Based Picture Captioning

By adding an image feature as an additive bias to the representation of the next anticipated word and by utilizing the image feature to gate the word representation matrix, the authors make the language model reliant on images in two separate ways. As a result, in the multimodal situation, the likelihood of producing a word wt is as follows:

$$P\,(W_t|W_1.........W_{t-1}, 1) \qquad (2)$$

where I is an aspect of the image. Their approach uses a deep convolutional neural network to represent pictures, and it implements joint image text feature learning by back propagating gradients through the multimodal neural network model. With the use of this model, an image caption can be generated word by word, with the conditional production of each word beingthe picture feature and previously generated words. By adapting a Recurrent Neural Network language model to multimodal instances for explicitly modelling the probability of creating a wordconditional on a given image and previously generated words, Mao et al. are able to provide innovative captions for images. Their method employs a deep convolutional neural network to extract visual data from images and a recurrent neural network with a multimodal component to simulate word distributions conditioned on image features and context words. Each unit in the Recurrent Neural Network language model is made up of an output layer y, a recurrent layer r, and an input word layer w. The computation carried out by these three layers is displayed at the last unit of the Recurrent Neural Network language model as follows:

$$x(t) = |w(t)r(t-1)| \qquad (3)$$

$$r(t) = f(U - x(t)) \qquad (4)$$

$$y(t) = g(v.rt) \qquad (5)$$

where U and V are matrices of weights that must be learned, and f() and g() are element-wise non-linear functions. The following equation is used by the multimodal component to calculate its layer activation vector m(t):

$$m(t) = g_m(V_w.w(t) + V_r.r(t) + V_{\boldsymbol{I}}.\boldsymbol{I}) \qquad (6)$$

where the function gm is non-linear. The image feature is I. The weight matrices Vw, Vr, andVI must be learned. By mapping and combining them, the multimodal component unifies picture features and scattered word representations. A perplexity-based cost function is minimized via back propagation to train the model. A strategy to align picture areas represented by a Convolutional Neural Network and phrase alignment is proposed by Schuster and Paliwal. A bidirectionalrecurrent neural network represents segments. A multimodal Recurrent Neural Network model can be learned from. Create sections of the image descriptions. After using their approach, employing related neural networks to represent picture regions and phrase segments, a structured objective is employed to map Bringing together visual and textual data in one location and linking each textual feature that describes the region to the region feature. The Then, aligned two modalities are used to train a multimodal system. Model of Recurrent Neural Networks, which can be utilized to predict the given a characteristic in an image, the likelihood of generating the next term words in context. It is well known that Recurrent Neural Networks struggle with gaining knowledge of long-term dependence. To remedy this Chen and Zitnick propose to dynamically create a visual representation of a picture as a caption to address the shortcomings in image captioning. For it, long-term graphic concepts are being created in order to be kept in mind throughout this approach. In order to accomplish this, a collection of latent variables Ut-1 are presented to codify visual interpretation. The number of created words Wt1 already. Utilizing these latent the likelihood of producing the word wt given various variables is as follows:

$$P\,(w_t.V\;|W_{t-1},\,U_{t-1})\;=\;P\,(w_t|V.W_{t-1}.U_{t-1}) \qquad (7)$$

where V stands for observable visual features and Wt-1 for created words (w1,..., wt-1). The authors implement the aforementioned concept by incorporating recurrent visual hidden layer u into the recurrent neural networks. The recurrent layer u is useful for both forecasting the following word wt and reconstructing the visual properties V from prior words Wt-1.

1.1.2 Captioning images using the Encoder-Decoder Framework

The encoder-decoder architecture is used to create captions for images. It was inspired by current developments in neural machine translation Fig. 2 depicts the general organization of encoder- decoder based image captioning techniques. Sentences from one language into another are initially translated using this framework. The argument that image captioning can be treated as a translation problem, where the input is an image and the output is a sentence, is motivated by the neural machine translation theory. In the framework's image captioning techniques, an encoder neural network first transforms an image into an intermediate representation, after which a decoder recurrent neural network uses the intermediate representation as input to create a sentenceword-by-word. In order to integrate joint image-text embedding models and multimodal neural language models, Kiros et al. introduce the encoder-decoder framework into picture captioning research. As a result, given an image input, a sentence output can be created word by word like language translation. They employ a deep convolutional neural network to encode visual data and long short-term memory (LSTM) recurrent neural networks to do so for textual data . Then, encoded visual data is projected onto an embedding space covered by LSTM hidden states that encode textual data by maximizing a pairwise ranking loss. A structure-content neural language model is employed in the embedding space to decode visual cues conditioned on context word feature vectors, enabling word-by-word sentence production. Vinyals et al. use a deep convolutional neural network as an encoder to encode images and long short-term memory (LSTM) recurrent neural networks to decode derived image attributes into phrases, both of which are inspired by neural machine translation. With the use of the aforementioned framework, the authors define picture captioning as estimating the likelihood of a sentence based on an input image:

$$S_{*} = arg_S max P\,(S|1\,:\,\Theta) \qquad (8)$$

Vinyals et al. model P(St | I, S0,..., St-1;) as a hidden state ht that may be modified by the update function below using a long short-term memory neural network:

$$h_{t+1} = f(h_t, x_t) \qquad (9)$$

where xt is the input to the neural network for long short-term memory. While xt is a characteristic of previously predicated context words in other units, it is an image feature in the first unit. By maximizing the likelihood of sentence picture pairs in the training set, the model parameter is obtained. Possible output word sequences can be predicted using the learned model using either sampling or beam search. Donahue et al. use a deep convolutional neural network for encoding and long short-term memory recurrent networks for decoding to provide a sentence description for an input image, similar to Vinyals' work. Donahue et al. offer both picture characteristics and context word information to the sequential model at each time step as opposed to just the first stage, which is the difference. Using the encoder-decoder framework to address the issue of image captioning has shown promising results. Encouraged by the success, strategies attempting to improve this framework are put forth in order to get improved performances.

2.RESEARCH METHODOLOGIES

Hybrid Model Both descriptive and analytical elements may be present in a model. To reason about a system, logical relationships in a descriptive model can be investigated and inferences made. The results of logical analysis, however, are very different from those of a quantitative chemical examination of system attributes. We first conducted a poll of people utilizing an onlineform creator and data collection service to acquire information regarding people's awareness.

3.PUBLIC SURVEY

For the purpose of collecting that will required for the research, a survey bot was deployed for thepurpose of collecting data from the public

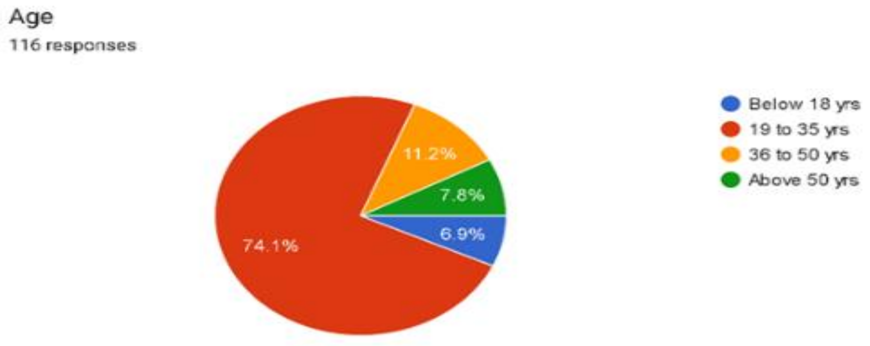
3.1Questioners

1. Do you read Captions written along the image
2. Does it help you to understand the context better
3. Are you aware of any program that can automate this process
4. Do you think a computerized program can automate the process of caption generation
5. If yes, will it be able to interpret the content on its own accurately
6. Do you know about such automation being implemented in any field
7. How do you believe this application for automatically creating image captions would be useful in day-to-day life
8. Do you believe it will function well with various state-specific languages

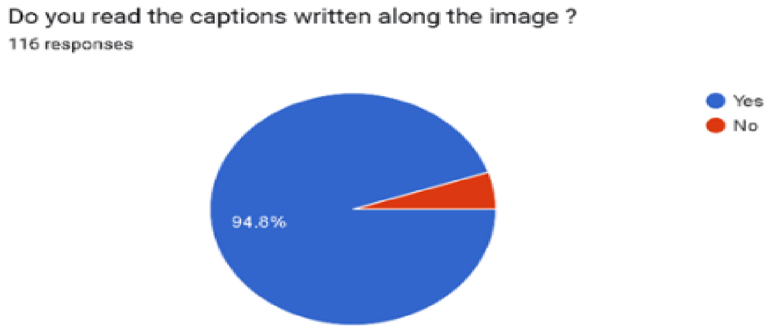
Do you think using an automatic image caption generator could lead to a gradual or significant decrease in work time

3.2Result

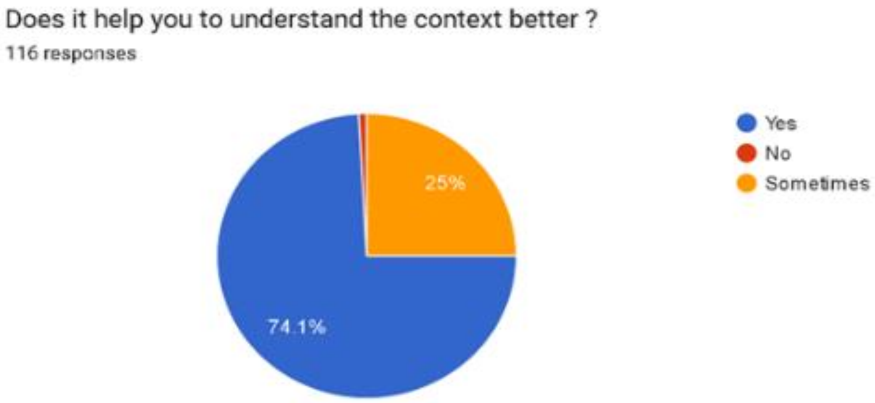
Most of the participants that participated in the survey were between the age range of 19-35.



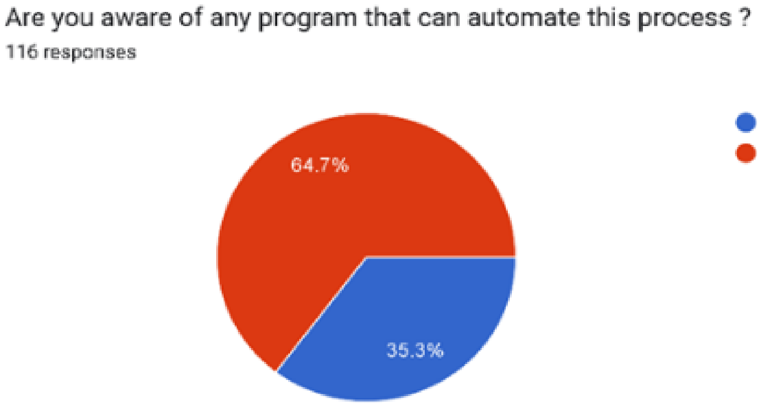
From the survey, it is observed that most of the people read the caption given along with theimage.



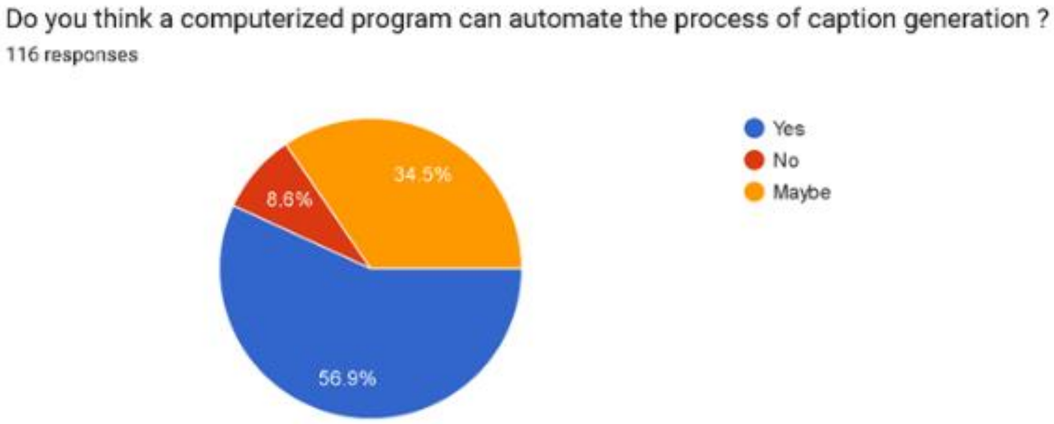
And for most people agree with the fact that the caption of an image helps them to understandthe context better.



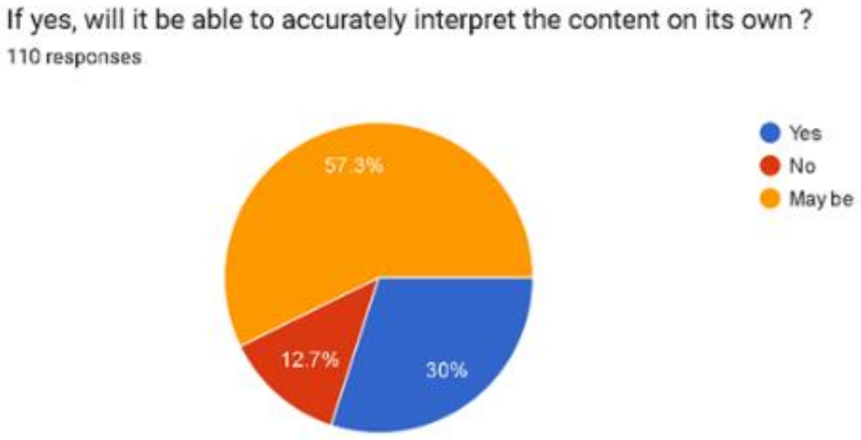
But it was observed that most people were not being aware of any software that was able to auto-mate the process of captioning.



On being asked whether a computerized program could be able to automate the process of captiongeneration for us more than half of the people gave us a positive response



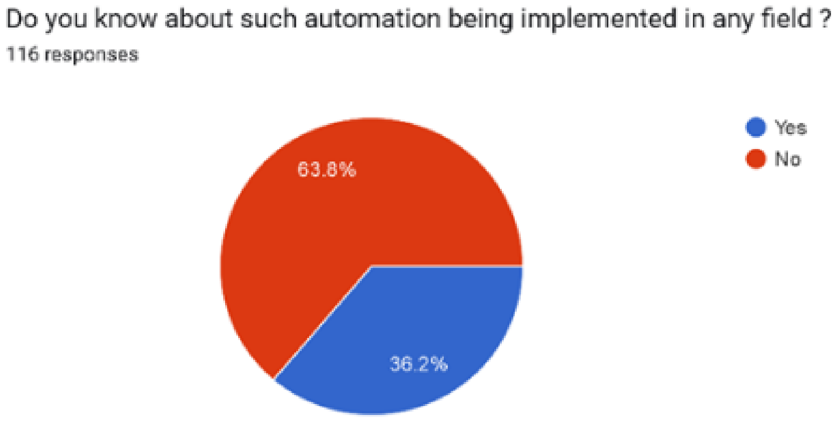
On being asked whether the computer program will be able to accurately interpret the contentmost of the people were not sure it would be able to.



If yes, will it be able to accurately interpret the content on its own ?

110 responses

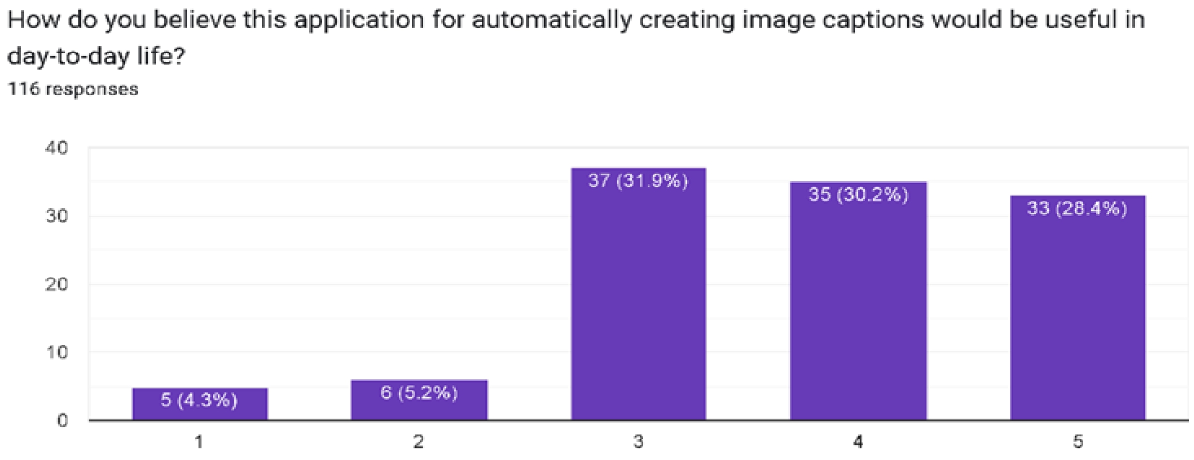
Most of the people were not aware of any field in which this kind of automation is being implemented



Do you know about such automation being implemented in any field ?

116 responses

Most people believe that the application for automation of the process would be helpful in day-to-day life



How do you believe this application for automatically creating image captions would be useful in day-to-day life?

116 responses

4.HYPOTHESIS TESTING

In order to draw conclusions about a population parameter or probability distribution, statistical reasoning known as hypothesis testing involves analyzing data from a sample. A hypothesis is first formulated with relation to the parameter or distribution. The shorthand for this is the null hypothesis, or H0. The null hypothesis is then contrasted with the alternative hypothesis (designated Ha), which is the complete opposite. The hypothesis-testing method decides if H0 can be rejected based on sample data. The alternative hypothesis Ha is valid if H0 is disproved, according to the statistical result.

For this paper,

Null hypothesis (H_0): Automation of caption generation is not the best way of captioning Alternative hypothesis (H_a): Automation of caption generation is the best way of captioning

4.1Test Statistics

There are three tests that can be used to decide whether or not the null hypothesis should be rejected. They are:

1. Chi-squared test
2. T-student test (T-test)
3. Fisher's Z test

A two-tailed T-student test will be used in this paper. When comparing the means of two groups that are connected in some way, a t-test is an inferential statistic that assesses whether there is a significant difference.

Level of Significance

The significance level is the likelihood that the null hypothesis will be rejected when ~~it~~ isconfirmed (also known as alpha or).

Level of Confidence

The confidence level shows the likelihood that a statistical parameter's position is correct (suchas the arithmetic mean) measured in a sample survey is also true for the entire population.

Sr.no	Data
1	94.5
2	73.6
3	35.5
4	56.4
5	29.5
6	34.5
7	53.6
8	86.4
Mean	58
Standard Deviation(s)	24.71217167

Level of significance = 0.05Level of confidence = 95

The number of standard deviations that separate a t-score (or t-value) from the t-mean distribution. The formula to find t-score is:

$$t = (\bar{x} - \mu)/(s/\sqrt{n}) \qquad (10)$$

where x is the sample mean,

μ is the hypothesized mean,

s is the sample standard deviation,and n is the sample size.

The p-value, also referred to as the probability value, expresses how likely it is that your data occurred under the null hypothesis. Finding the equivalent p-value is possible once we are awareof the value of t. The null hypothesis can be rejected and Automation of caption generation is thebest way of captioning if the p-value is less than a certain alpha level (popular choices are .01, .05,and .10).

Calculating t-value:

Step 1: Identify the alternative and null hypotheses.

Null hypothesis (H0): Automation of caption generation is not the best way of captioning. Alternative hypothesis (Ha): Automation of caption generation is the best way of captioning.

Step 2: Find the test statistic.

The postulated mean value in this situation is taken to be 0.

$$t = (\bar{x} - \mu)/(s/\sqrt{n}) = (58 - 0)/(24.71217167/\sqrt{6}) = 6.638 t - value = 6.638 \qquad (11)$$

Calculating p-value:

Step 3: Calculate the test statistic's p-value.

The p-value is computed using the t-Distribution table with n-1 degrees of freedom. The sample size for this study is n = 8, hence n-1 = 7. It provides a p-value when the observed value is entered into the calculator. In this case, the p-value returned is 0.00029366.

We can reject H_0 at the significance level of 0.05 because your p-value does not exceed 0.05.

Therefore, we have enough information to conclude

Automation of caption generation is the bestway of captioning.

5.FUTURE DIRECTIONS FOR RESEARCH

Considering how new the issue of automatically captioning images is, significant advancement has been made as a result of the work of scholars in this area. The effectiveness of image captioning could certainly use some improvement, in our opinion. First, given the rapid advancement of deep neural networks, applying more potent network structures as language and/or visual models willunquestionably boost the effectiveness of image description creation. Second, since captions for photos are simply word sequences whereas images are made up of objects scattered across space,it is crucial to look at the presence and hierarchy of visual concepts in captions. The effectiveuse of the attention mechanism to create picture captions will also remain a key study area since this problem suits the attention mechanism well and because it is proposed to perform a variety of AI-related activities [129]. Third, research on using unsupervised data, such as from photos alone or text alone, to improve image captioning will be promising due to the lack of coupled image-sentence training set. Fourth, present methods generally concentrate on creating general captions describing the contents of images. However, as noted by Johnson et al. [130], picture description needs to be firmly rooted in the aspects of the photographs in order to be relatable tohumans and useful in real-life settings. As a result, one of the future study topics will be image captioning founded on image regions. Fifth, while task-specific image captioning is required in some situations, the majority of existing approaches are geared to provide image captioning for generic cases. It will also be fascinating to conduct research on several unique scenarios where image captioning issues arise.

6.CONCLUSION

We give a survey on image captioning in this paper. We categorize picture captioning methods into many groups based on the strategy used in each method. The strengths and weaknesses of each sort of job are discussed along with representative approaches from each area. We start out by talkingabout early image captioning research, which is primarily retrieval- and template-based. Next, neural network-based techniques are the main focus of our research because they produce cutting- edge outcomes. We then separated them into subgroups and examined each subcategory separately because different frameworks are employed in neural network-based methodologies. Following that, benchmark data sets are used to compare state-of-the-art approaches. Finally, we outline potential future avenues for automatic picture captioning research.

REFERENCE

[1] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4259–4267, 2015.

[4] Girish Kulkarni Visruth Premraj Sagnik Dhar, Siming Li, Yejin Choi Alexander C Berg Tamara, and L Berg. Baby talk: Understanding and generating simple image descriptions. 2013.

[5] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

[6] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance ofa real-world scene? *Journal of vision*, 7(1):10–10, 2007.

[7] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detec- tion with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[8] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source do- main generalization. In *Proceedings of the IEEE conference on computer vision and pattern*

recognition, pages 87–97, 2016.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[10] Ankush Gupta, Yashaswi Verma, and C Jawahar. Choosing linguistics over vision to describe images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 606–612, 2012.

[11] Patrick Héde, Pierre-Alain Moëllic, Joël Bourgeoys, Magali Joint, and Corinne Thomas. Auto- matic generation of natural language descriptions for images. In *Coupling approaches, coupling media and coupling languages for information retrieval*, pages 306–313. 2004.

[12] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a rankingtask: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[13] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[17] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*, pages 3177–3184. IEEE, 2011.

[18] Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image cap- tioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, 2014.

[19] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[20] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Trans- actions of the Association for Computational Linguistics*, 2:207–218, 2014.

[21] Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Efficient image annotation for automatic sentence generation. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 549–558, 2012.

[22] Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 444–454, 2011.

[23] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.