IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 6, Issue 2, 2018

Design of a Picture-seeing and Talking SystemBased on Attention Mechanism

K.VAMSHI KRISHNA[1], K.SRINIVAS[2], CH.SUPRIYA[3], V.SWARUPA[4]

Abstract This research presents a deep loop picture title generating approach to address this issue. For a simple physical picture, this approach leverages recent advances in computer vision and machine translation to provide descriptive phrases in plain language. When given a training picture, the model is optimized for producing a target description phrase. The MSCOCO data collection was essential in filling out the training set, and in the fine-tuning phase, I included in a few feature images that I had expressly sought for online. The model's capacity to accurately describe images at a fundamental level was validated by experimental testing on many datasets. I've shown subjectively and numerically that this model is more accurate in the case of straightforward physical images. The finished product accepts a qualifying picture as input and produces a natural language statement describing the image's key substance.

Keywords: *Using MSCOCO data for AI, deep learning, computer vision, and machine translation*

1. Introduction

Research Background

As its name indicates, "image captioning" (image captioning) entails producing a phrase or group of keywords describing a picture based on its input, with the input sentence needing to be grammatically correct and the tense correct. Children as young as kindergarten started learning to read and write by focusing on images and chatting about them [1]. This is a demanding and difficult assignment for computers, though. Because we want the computer to be able to look at pictures and have conversations, it needs to be able to understand the image by identifying objects and scenes within it, and it also needs to be able to organize language in a way that is comparable to that of a human, using a natural language to describe what it sees. Below is an illustration of three photos and their respective explanations. Reading and speaking are seen as translation tasks in this study; however, the more commonplace translation from language to language is replaced with the less familiar translation from picture to language. When discussing translation, it is impossible to avoid mentioning the use of machine translation (Machine Translation) programs, which take a phrase in one language and provide a literal translation into another. The latter may be a literal translation into another language, or it may be a condensed version of the former using a set of essential terms, known as an abstract. The so-called "collision produces sparks" principle means that the field of image-to-language translation may learn from the field of language-to-language translation in machine translation, and this means that many studies on picture caption creation can be categorized as machine translation studies in the subject of natural language processing. CV is seeing growth and development as a field. As a result, the study of visuals and verbal communication based on the attention model progressively took shape. Nearly all machine translation algorithms have previously been constructed using mathematical and statistical techniques before the advent of deep learning. As the discipline of natural language processing (NLP), and natural machine translation specifically, has progressed, so too has the adoption of models based on deep learning technology [1]. In the area of natural language processing (NLP), machine translation has been dominated by the "Sequence to sequence" (Seq2Seq) model built on deep learning networks since 2014. With the introduction of the attention mechanism in 2015, the Seq2Seq model was able to fix many of the fundamental flaws in machine translation.

ASSOCIATE PROFESSOR[1,2,3,4]
ELECTRICAL AND ELECTRONICS ENGINEERING
TRINITY COLLEGE OF ENGINEERING AND TECHNOLOGY, PEDDAPALLY
(vamshi.komuravelli@gmail.com),(klsrinivas.tcek@gmail.com), (supriyajashu21@gmail.com), (velpulaswarupa@gmail.com)

to new contexts.

Research Significance

Build a model for attention-based image caption creation such that the computer can create a caption that accurately describes the contents of a given picture. The core challenge of automatic caption creation is adapting from sight to speech, with a very basic explanation: just look at the image and describe what you see. In the same way that kindergarten instructors want their students to look at images and explain what they see, the authors of this research expect the algorithm they developed to provide a natural-sounding statement

describing the input image.

The issue of automatically finding the matching descriptive text after a given picture is addressed by image description. This very difficult artificial intelligence research subject lies at the crossroads of computer vision, natural language processing, and machine learning. It has a wide range of possible uses, including but not limited to semi-supervised, unsupervised, transfer, representation, and few-shot learning; identification; detection; segmentation; and posture estimation. It's a major obstacle to overcome in the study of AI, and the results of this investigation will play a crucial directing and foundational role in AI's future evolution.

Research Status at Home and Abroad

Since the inception of deep learning, several organizations and people have been drawn to study its applications. New concepts and works related to the future of deep learning have been emerging at a fast pace in recent years, much like mushrooms after a rain. Many of the world's largest corporations are also engaged in deep learning research. In addition to theoretical work, tech giants like Apple, Microsoft, Google, Amazon, etc. have produced several useful products like the intelligent voice robot "SIRI" from Apple and "Xiaoice" from Microsoft, as well as the world-famous first player in the board game Go, "Alpha Dog" [2].

The United States may be behind the rest of the world when it comes to deep learning research, but the country is quickly catching up. There is a steady influx of major corporations into the market. Alibaba, Tencent, Baidu, and Huawei are just a few examples of homegrown businesses that have contributed to this area of study and made significant strides in doing so. Some smaller firms are also doing research in this area with the aim of applying their findings to other sectors.

The Main Research Content of This Topic

The end objective of this subject is to develop a fully trained image caption generation system that can generate a natural language phrase to represent the major content of a given input picture. This necessitates the following research:

Learn all there is to know about deep learning, from the basics of how neural network models function to advanced techniques for applying what you've learned. Get up to speed with Python and the Pycharm development environment.

The model's capacity to describe images is continually improved by tweaking the code or undergoing specialized training, and the process begins with using the data set to train a large number of the intended model.

2. Research Ideas and Tools

Research Ideas

Generating natural language descriptions from visual data has long been a popular research area in computer vision, although most existing solutions focus on video. Combining visual primitive recognizers with structured formal languages results in a fairly sophisticated system. For example, And-Or diagrams and logical systems may be translated into "human languages" using rule-based translation tools. Most of these systems are intentionally created; their capacity to maintain equilibrium is weak; and their applications are restricted to relatively narrow domains like surveillance and sports [3].

Another recent focus has been on the challenge of automatically creating text captions for images based on their natural language descriptions. Recent developments in object, property, and location recognition power natural language generation systems, but their expressiveness is severely constrained and they always have major problems.

The task of determining how best to order available picture descriptions has been mostly resolved at this point in time. The aim behind this method is to simultaneously capture both text and visuals in the same vector space. Just get the output description that looks like the input picture. Instead of attempting to construct extensive descriptions, neural networks more typically incorporate pictures and phrases to break the visual information into multiple feature values [3]. The aforementioned techniques, even when applied to objects that have been taught using the training data, are not able to accurately characterize the composition of unseen objects. In addition, there is no method for gauging the accuracy of the produced descriptions.

Thus, in this paper's architecture, we mix a deep convolutional network CNN for image classification with a recurrent network RNN for ranking models to provide a dedicated network for producing picture captions. This SISO network setting is used to hone the RNN's training. Instead of a text, this system takes in a picture that is then analyzed by a convolutional network. Similar models have been developed in the past using neural networks with a feed-forward network to foretell the visual representation of the next and the preceding words. The identical job of prediction is then performed using a recurrent neural network. Similar ideas have been proposed before, but this study takes them a step further by using a more robust RNN model and directly providing the RNN's visual input model, so enabling the RNN to interpret the objects it

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 6, Issue 2, 2018

tracks as text. This distinction is what allows our system to perform better in experiments. The combined multimodal embedding space [2] is built utilizing robust computer vision models and LSTMs. Figure 1 shows how this is done.

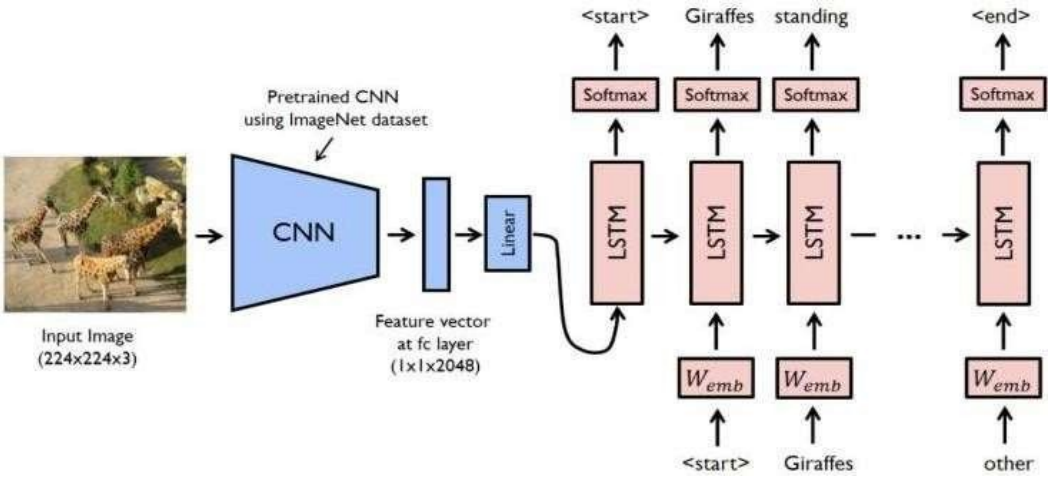


Figure 1. Schematic diagram of the overall model

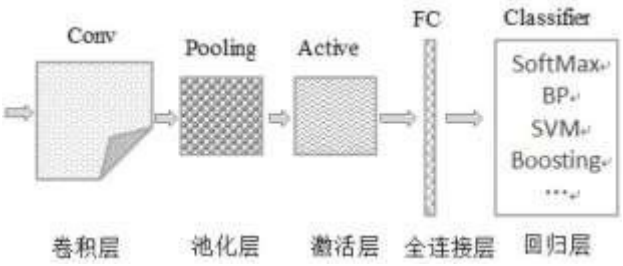
International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 6, Issue 2, 2018

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) form the basis of this NIC model's neural network. As seen above, it can generate a whole phrase in natural language that accurately conveys the meaning of the graph extracted from the input picture.

Basic Introduction to Using Model Database

CNN (Convolutional Neural Network)

Convolutional neural networks (CNNs) were first developed between the years 1880 and 1990. It finds particular success in the areas of computer vision and natural language processing, among others. In addition to being one of the foundational algorithms of deep learning, a feedforward neural network with convolutional computation and a deep structure is also a part of this model. Among other names, "Translation Invariant Artificial Neural Networks" [5] describes what we call a Convolutional Neural Network. As a matter of fact, this network is now considered a major advancement in computer vision.



There are several kinds of neural networks, and one of them is the convolutional neural network. It differs from the standard neural network in that its convolution operation takes the place of matrix multiplication. The convolution process in a convolutional neural network [4] allows for the detection of the planar structure that makes efficient use of the input data. Therefore, as compared to conventional neural networks, convolutional neural networks perform much better in the areas of image and voice recognition.

The field of Deep Learning has seen a recent surge in interest in Convolutional Neural Networks. This phenomenon occurs because it bypasses the need for time-consuming and complicated preparation by processing the raw data from the images being entered.

process. The term "convolution" is at the heart of what it is called. Features like BP and HOG from traditional machine deep learning may be thought of as a subset of convolution. The term "convolution" refers to the usage of a number of This technique provides the most accurate description of the image's abstract characteristic.

Figure 2. Working principle diagram of CNN operation

Convolutional neural networks, like the one seen above, employ a sliding window (convolution kernel) to efficiently "screen" the input image area, multiply the associated pixels one at a time, and then collect (I*K), the pixel's intensity value. The newly-published convolution results are available now. Since convolution is sometimes compared to a sieve, and since it employs a convolution kernel to "multiply and accumulate" the original picture twice (which is an integral), the following function is used to represent it in this article. A formula:

$$y(t)^{\infty} = \int_{-\infty} x(p)\,h(t-p)dp = x(t)*h(t) \quad (1.1)$$

Let's take a look at a typical CNN example, for the image processing process with a resolution of 28*28:

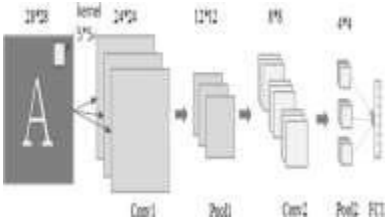
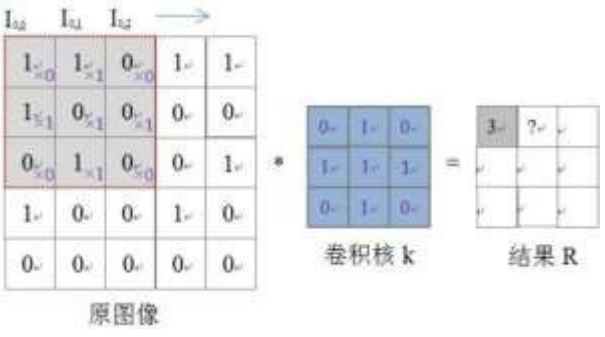


Figure 3. Schematic diagram of image processing by CNN

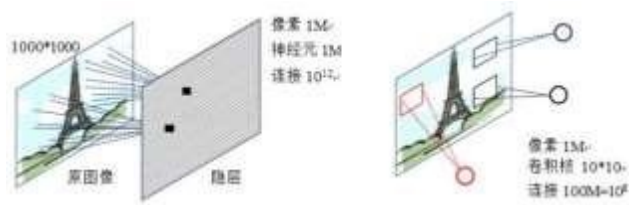
The convolution layer seen in the image (with a 5*5 convolution kernel and a Step value of 1) is responsible for extracting features from the input data. The pooling layer, also known as downsampling (Sample), is responsible for reducing complicated high-dimensional data; the fully connected (FC) layer uses the pooling layer's data directly in its calculations. Hidden layers are a collective name for the several intermediate layers used throughout this work. Next, learn more about CNN by diving into the details of what each layer does.

Figure 4. Schematic diagram of each layer of CNN

1. Convolutional layer and weight sharing



International Journal of
Information Technology & Computer Engineering
ISSN 2347–3657
Volume 6, Issue 2, 2018



In order to extract characteristics from data, the neurons in a live body must adhere to a strict set of rules. As can be seen in the graphic below, the mathematical approach for processing an image with 150 pixels is quite complicated since it requires 1012 connections, which results in a large number of weights..

Figure 5. Schematic diagram of the principle of CNN processing theweight problem

2. How can I get an answer to a multitude of weight calculation puzzles? The term "classification" is crucial since it indicates that different neuron types have different weights. To mimic the way neurons process information, this article uses simply a convolution operation with 10,000 weights, as illustrated in the image above. In particular, the efficient convolution that occurs as part of this process requires no complex parameter operations from the computer. A more straightforward and comprehensive technique of computation is used.

3. **Activation layer**

4. The activation layer's neural network artwork accurately depicts how neurons process information. If you utilize the now-common ReLU function in CNN, you may forget about feature "scatter" and "pre-training" altogether since you've successfully solved the issue of gradient diffusion. There is no hard and fast rule on when to add activation layers following convolutional or pooling layers. The activation layer is often omitted when constructing a basic neural network.

5. **Dropout layer**

The problem of "overfitting" in neural networks was another major issue until the Dropout layer was introduced. Traditional approaches to addressing the "over-fitting" problem involve model averaging—that is, avoiding it by training multiple networks for weighted combination—which is analogous to the under-fitting that led to the "gradient diffusion" problem but requires more computation. Eventually, issues will arise [5]. The Dropout technique is an excellent tool for dealing with any issue. Hinton was the first to suggest it. Neurons only remember the weights' parameters and update them during the next training session. In addition, each training uses a unique set of neurons selected using a predetermined random approach, allowing the neuron nodes to take turns firing. The human brain more closely resembles this random process.

Figure 6. Schematic diagram of the working principle of trainingneurons

Each Dropout procedure is analogous to a simplification of the aforementioned network. Each node in the simplified network may take part in the weight update, and Dropout is executed numerous times during the training phase to ensure that every node is learning. And training is a straightforward procedure that has the potential to provide outstanding outcomes.

Dropout is an effective and intuitive strategy that was developed from a commonsense knowledge of biology and has been shown by several trials, albeit its mathematical underpinnings may be lacking.

For what reason does it converge so well? How can one's theory prevent them from becoming mired in local maxima? While it's possible that just knowing about the issues involved isn't enough to provide satisfactory solutions, there are occasions when such effort is unnecessary. Don't stress too much over the intricacies of the reasoning.

Layer 6: All connections made

The completely linked layer may be grasped conceptually as a streamlined data computation. In the end, even deleted.

Layer 7: Regression

Although the regression layer stands on its own, it is often considered to be a component of the fully connected layer and serves just to link the processes in this area. The idea of regression is not hard to grasp. Types of regression include those described above, such as logistic regression and Gaussian regression. It's just a generic term for a method of organizing images by their many characteristics. Its primary function is to convert between P and K dimensional vectors, and its formula is as follows:

RNN model has an input layer, a hidden layer, and an output layer, as indicated in the picture below:

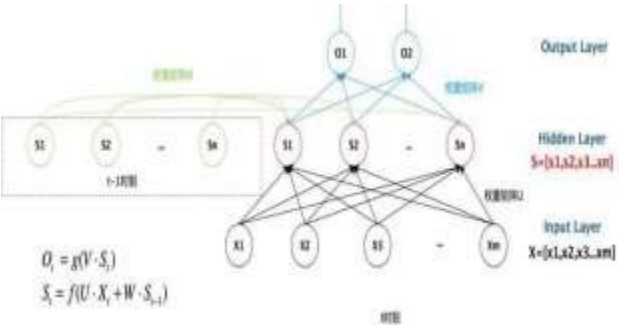
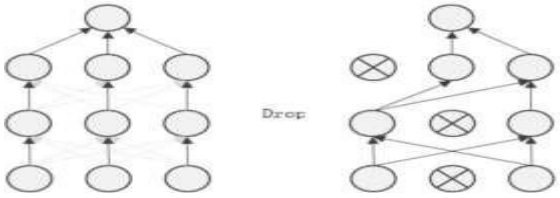


Figure 7. Schematic diagram of a simple RNN model

Thereby, the probability represented by the correspondingcategory is obtained, which is the classification result that this article wants.

RNN (Recurrent Neural Network)

RNN is a kind of neural network with a unique structure. The output of an RNN differs from that of a DNN or a CNN in that it not only reflects the input from the preceding time period, but also provides the network with a "memory" function based on the



International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 6, Issue 2, 2018

information that has already been processed. When an RNN generates an output, it does so only after giving careful "consideration" to both the input and all prior outputs [7]. The network's specific working principle is that it will remember the past data and incorporate it into the present calculation of output; this is accomplished by connecting previously "isolated" nodes between hidden layers, so that the output of the hidden layer incorporates not only the output of the input layer, but also the output of the hidden layer from the previous instant [8].

If RNN (Recurrent Neural Network) isn't necessary, why do we need it? Regular neural networks, as we've established, have a hard time dealing with more than one input at a time, and the results of one input have nothing to do with the results of another. However, the RNN's sequence processing capabilities become crucial when dealing with jobs that need understanding the connection between one input and the next. First, let's take a look at a

Figure analysis reveals that x is a vector group representing the input layer's value, s is a vector group representing the hidden layer's value, U is a weight matrix connecting the input and hidden layers, and o is a vector group representing the output layer's value, with the corresponding V being the weight matrix connecting the hidden and output layers [9].

The graphic shows that the recurrent neural network's hidden layer value s is dependent not only on the current input x but also on the value s of the preceding hidden layer. As the proportional value of the current input, this time around's weight is stored in the weight matrix W [10]. Here are the relevant detailed diagrams:

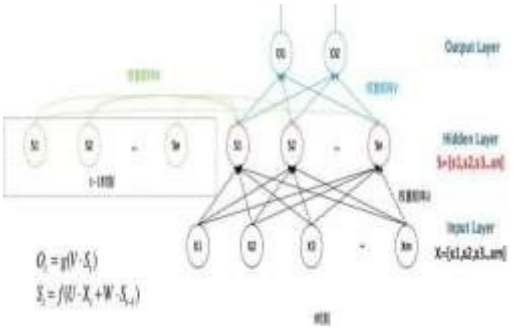
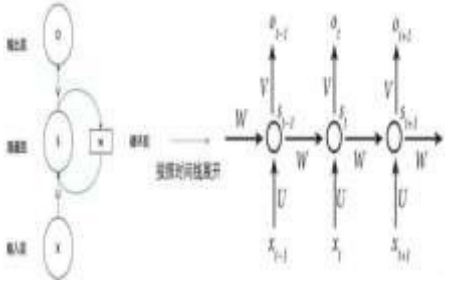


Figure 8. Schematic



As can be seen from the above figure, the hidden layer of the previous time step can affect the hidden layer of thecurrent time step.

Expand the above picture to get the following schematic expansion diagram:

Figure 9. Schematic diagram of RNN unfolding according to the timeline

his looks very clear. The following formula can be used to specifically express the calculation method of the recurrentneural network:

$$O_t = g\left(V \cdot S_t\right),\ S_t = f\left(U \cdot X_t + W \cdot S_{t-1}\right)$$

(2-3)

Finally, an overview of the RNN is given:

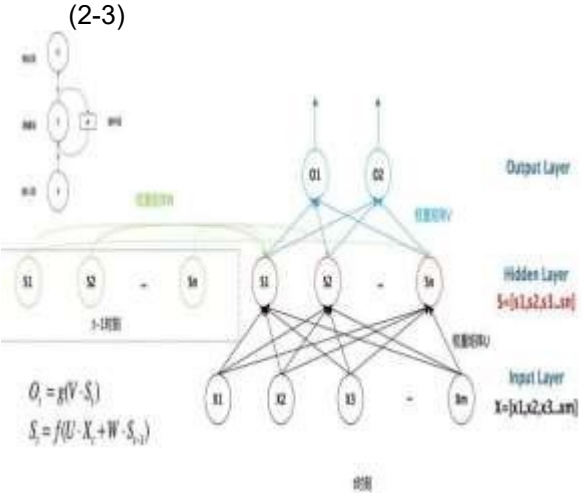


Figure 10. RNN overview diagram

data collection contains more than 1.5 million records. The primary benefits

This research uses the MSCOCO data set for its primary training and testing data set after careful evaluation and comparative analysis among many parties. Because of its size and depth, the COCO data collection is the primary focus of this article. The goal of this dataset is scene comprehension, and it does so by intercepting photos of complicated everyday scenarios and precisely calibrating the targets inside them. There are 91 categories of things included in the photographs, with 328,000 pictures and 2.5 million labels [11]. With over 80 different categories and over 330,000 photos, COCO is currently the biggest semantic segmentation dataset available online. There are a total of 300,000 labeled images in this dataset, and its primary strengths lie in three areas: wide-ranging detection targets, accurate differentiation of contextual connections between targets, and pinpoint localisation of target features' two-dimensional coordinates. Here are some of its key characteristics:

ISSN 2347–3657
Volume 6, Issue 2, 2018

1) Identifying Your Audience
2) using optical character recognition
3 Multiple targets in a single picture
4) There are almost 300,000 pictures
5) More than 2 million occurrences
Sixty-eight Classes
7) Images often have 5 objectives.
8) There are essentials for a hundred thousand people.
Below is an illustration of some MSCOCO data.

Figure 11. Example of MSCOCO data

In all, two versions of the COCO dataset have been made available to the public. 2014 saw the debut of the product. There are 270,000 segmented portrait photos and 886,000 segmented object images, as well as 82,783 training samples, 40,504 validation samples, and 40,775 test samples. There were a total of 165,482 training samples, 81,208 validation samples, and 82,434 test samples available in the second release of 2015. There are a few crucial points to remember while working with this dataset:
1. the image's name, indicated by the #pointed-to string file_name;
2. dimensions height and width #The dimensions of the picture that was pointed to;

Third, the image's own label is indicated by the id #. There is no pattern to the numbers. Like the serial number on an ID card, it may be thought of as information inherent to the picture.
Annotation 4 Indicates a list with many dictionaries whose titles are all links to images with annotations. Images from the coco dataset with annotated examples:

Figure 12 An example of annotated images in the COCO dataset

Environment Construction and Libraries

PyCharm

To increase efficiency while working with the Python programming language, many developers nowadays turn to PyCharm.The IDE also incorporates a number of additional cutting-edge features, and skilled web developers have found that using Pycharm has led to some pleasantly surprising outcomes.

Visit https://www.jetbrains.com/pycharm/download/ to get Pycharm. After launching, users may choose their preferred OS (Windows, macOS, or Linux) and thereafter download either the free community edition or a paid professional version.

Anaconda and Libraries

In addition to Python itself, the Anaconda installation also contains the Python package management conda, as well as a number of scientific computing-related programs. The advantages of using Anaconda are as follows:

1) Do away with platform-specific issues and address underlying dependencies. When you install a package, any issues that arise as a result of its

ISSN 2347–3657
Volume 6, Issue 2, 2018

dependencies are taken care of automatically.

2) The idea of a "virtual environment" exists. Each environment may have its own Python version and collection of packages, and they are kept separate from one another. It is compatible with the system and may be turned on and off at will. It's also tidied up if you decide to erase it.

This content was taken straight from the source. The library may then be managed and an appropriate environment built using conda after the download is complete. To write this system code, you'll need to install the necessary environment toolkits:

Tensorflow

TensorFlow is a software library used for high-performance numerical calculations that includes multiple open-source components. Users may use it to deliver computing workloads across a wide range of platforms and devices because to its adaptable architectural architecture.

Since a data flow graph is used in TensorFlow's processing, we will first construct one in this post and then insert the tensor data into it. Graph nodes are mathematical processes, while edge connections are multi-dimensional data arrays, or tensors [13]. During model training, tensors move constantly from one node to another in the dataflow graph. This is the inspiration for the term "TensorFlow." In machine learning, there are typically four different kinds of values:

(1) scalar (scalar): a numerical number, which is the lowest running unit in the computation process,

TensorFlow's benefits may be seen most clearly in the following areas:

(1) TensorFlow's "tensor flow" design is particularly user-friendly. All tensor flow connections are shown clearly and sensibly for the user.

Two: TensorFlow is well-suited for distributed computing and can be easily deployed on CPU/GPU.

Thirdly, TensorFlow may be used with almost any platform. Not only does TensorFlow work effectively on Linux, Mac, and Windows computers, but it also functions well on mobile devices. The low-level nature of the code in TensorFlow and the extensive amount of user-generated documentation are two of the library's key drawbacks. In addition, users sometimes have to "rebuild the wheel" [13] for comparable and repetitive tasks. Although "the flaws do not hide the flaws," TensorFlow's extensive technical features and reliable performance have kept it at the forefront of various deep learning frameworks' adoption rates for some years.

such as "9" or "567".

A vector is an array of scalars in one dimension, such [1, 3.2, 4.6], etc.

Matrix (three) — A two-dimensional array of scalars.

(4) Tensor: A set of data that consists of arrays of several dimensions and may be interpreted as a high-dimensional matrix. Here is a look at Tensorflow's data flow diagram:

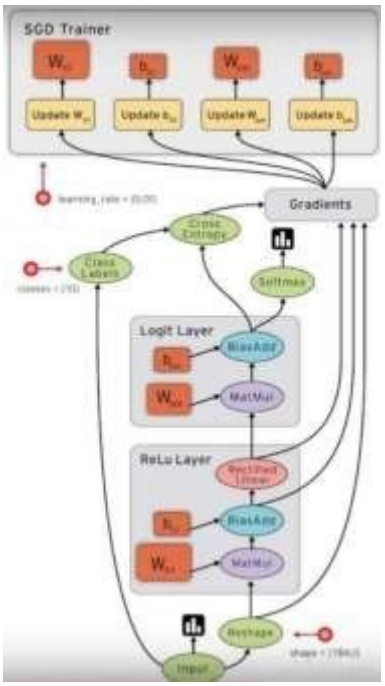


Figure 13. Tensorflow data flow diagram

NumPy

NumPy is a Python package for fast and precise numerical computing and analysis.
NumPy's primary features include:
To begin, there is the memory-conserving data structure ndarray.

Second, make data operations quick and avoid looping.
Third, software for accessing and editing files in memory mappings and on disk.
Replace the target's Fourier coefficients with newly generated linear functions and pseudo-random numbers, and then apply the Fourier transform.
Tools for combining C and C++ with other programming languages 5.
The rank of a NumPy array indicates the number of axes along which data may be sorted. Each dimension of a NumPy array is referred to as an axis. The components of an ndarray are as follows:
One such data location indicator is a "pointer."
2. The dtype information that represents the array's contents.
3. A tuple describing the form of the array.
Subscripting allows us to get the total number of students in a class together with their names, Chinese, math, and

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 6, Issue 2, 2018

English scores, but this is a time-consuming and inefficient method. In order to utilize Numpy for operation implementation, a custom data structure must be defined using the dtype keyword.

Figure 14. Download Tensorflow using conda

```
import numpy as np

studenttype = np.dtype({
    'names': ['name', 'chinese', 'math', 'english'],
    'formats': ['S32', 'i', 'i', 'i']
})
```

Figure 15. Use dtpye to define data structure code diagram

Then, when using array to define an array of real data, define the stype element attribute as the custom data structure above, so that the custom data structure can be called.

```
students = np.array([('zhangsan', 85, 72, 56), ('lisi', 88, 90, 68),
                     ('wangwu', 78, 66, 88)], dtype=studenttype)
```

Figure 16. Use array to define data structure code diagram

```
name = students[:]['name']
chinese = students[:]['chinese']
math = students[:]['math']
english = students[:]['english']
```

Figure 17. Code diagram for taking out demand value

Then take out all the values you need, here this article takes out all the values.

After the data is extracted, the data can be processed,for example, the average of the scores of the three students in each subject is required. In the Numpy library, mean()is used to find the mean.

```
print(np.mean(chinese))
print(np.mean(math))
print(np.mean(english))

# 结果
```

Figure 18. Output average code with mean

In addition, you can also perform addition, subtraction, multiplication and division operations on arrays, and remainder operations. An example of an array createdwith the two functions above.

```
import numpy as np
b = np.linspace(1, 7, 4)
c = np.arange(1, 8, 2)

print(np.add(b, c))          # 加法运算
print(np.subtract(b, c))     # 减法运算
print(np.multiply(b, c))     # 乘法运算
print(np.divide(b, c))       # 除法运算
print(np.mod(b, c))          # 取余运算

# 结果
[ 2.  6. 10. 14.]
[0. 0. 0. 0.]
[ 1.  9. 25. 49.]
[1. 1. 1. 1.]
[0. 0. 0. 0.]
```

Figure 19. Schematic diagram of addition, subtraction, multiplicationand division operation code

Calculate the maximum, minimum, mean, standard deviation, and variance in an array.

```
a = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
print(a.max())      # 数组中最大值
print(a.min())      # 数组中最小值
print(a.mean())     # 数组中平均值
print(a.std())      # 数组中标准差
print(a.var())      # 数组中方差

# 结果
9
1
5.0
2.581988897471611
6.666666666666667
```

ISSN 2347–3657
Volume 6, Issue 2, 2018

Figure 20. Schematic diagram of code for calculating the maximum andminimum values of an array

The above is a brief introduction to the functions and usage of CumPY.

OpenCV

OpenCV was established by Intel in 1999. After severalyears of development and optimization, it has gradually improved. As a research tool in the field of computer vision, it has a wide range of applications and is deeply favored by people. In this paper, it is mainly used for image processing and machine vision recognition. Thespecific principles are introduced as follows:

1. The following two more classic recognition of different image forms

Only black and white grayscale images are single-channel, as shown in Figure 21 below. Each pixel block corresponds to a value between 0 and 255 in the matrix. This value represents the grayscale of this pixel, from pureblack 0 to Pure white 255 [14]:

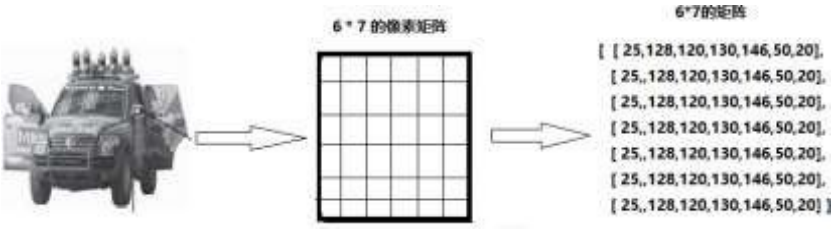


Figure 21. Schematic diagram of the storage form of black and white pictures

For color images, only three-channel images in RGB format are currently supported. The feature is that each pixel block is composed of different depths of three colors of red, green and blue. One pixel block corresponds to a vector in the matrix, such as [155,147,220], respectively representing the proportion of the three colors in this pixel block, as shown in the following figure [14]:

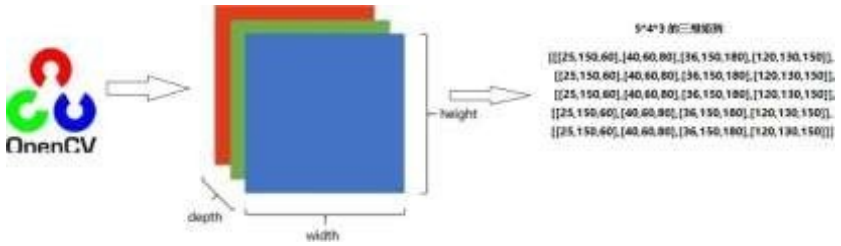


Figure 22. Schematic diagram of RGB image storage form

2. Image input and recognition (code and its introduction)

1) imread

imread(img_path, flag) #Read, return the command ofthe image

img_path #The path of the image, if the path is wrong,the image will be returned as none

flag: cv2.IMREAD_COLOR #Read color images, but do not recognize image transparency, all set to the defaultvalue cv2.IMREAD_GRAYSCALE #Read grayscale images, you can pass in 0

cv2.IMREAD_UNCHANGED #Read the image,including the alpha channel, you can pass in -1

2) imwrite

imwrite(img_path_name, img)

img_path_name: the name of the saved fileimg: file object

3. Image reading and writing (code and its introduction)

1) Pixel value

acquisitionimg = cv2.imread(r"C:\Users\Administrator\Desktop\roi.jpg") # get and set

pixel = img[100, 100] #[57 63 68], get the pixel value at (100, 100)

img[100, 100]=[57, 63, 99] #Set the pixel value

b = img[100, 100, 0] #57, get the pixel value of the bluechannel at (100, 100)

g = img[100, 100, 1] #63

r = img[100, 100, 2] #68

r = img[100, 100, 2]=99 #Set the red channel value# get and set

piexl = img.item(100, 100, 2)

img.itemset((100, 100, 2), 99)

2) The nature of the pictureimport cv2

img = cv2.imread(r"C:\Users\Administrator\Desktop\roi.jpg")

#rows, cols, channels

img.shape# returns (280, 450, 3), width 280 (rows), length 450 (cols), 3 channels

(channels)#size

img.size#returns 378000, the number of all pixels, =280*450*3#type

img.dtype#dtype('uint8')

3) ROI interception (Range of Interest)#ROI, Range of instrest

roi = img[100:200, 300:400]#Intercept lines 100 to 200,and the columns are the entire area of columns 300 to 400

img[50:150, 200:300] = roi#Move the intercepted roi tothis area (rows 50-100, columns 200-300)

b = img[:,:,0]#Intercept the entire blue channel

```
b, g, r = cv2.split(img)#Intercept three channels, whichis time-consuming
img = cv2.merge((b, g, r))
4) Add border
(padding)
cv2.copyMakeBor
der()
```

The above is the basic image processing process of opencv. During the download process, due to the temporary stop service of Tsinghuayuan, it cannot be downloaded directly in conda, so this article uses pip to successfully download opencv.

Summary of This Chapter

This chapter primarily presents the fundamental concepts and functionalities of various necessary models and tools, as well as the primary research ideas of this system's design. Knowledge of the particular environment and data packages is also introduced, as is information on how to obtain and apply different data sets and environments. Like fish need water to survive, you can't design without first learning the fundamentals, and without a fully configured environment and data package, you won't be able to use many of the tools at your disposal, which means you won't be able to write. The training and testing of the model both reveal the significance of the data set. The project necessitates a model of the system that has been trained to perceive images and understand speech.

3. Model Research

Model Design

In order to produce descriptions from photos, the authors of this research suggest a neural and probabilistic framework. Recent developments in statistical machine translation have demonstrated that, given a strong sequence model, optimal results can be achieved by directly improving the translation accuracy of input sentences in a "point-to-point" fashion. This model can be used not only for training, but also has a specific inference function. To "decode" the input feature encoding into the target phrase, this model employs a regression approach to transform it into a fixed collection of multidimensional vectors [16]. Therefore, the model will use the aforementioned technique to convey the content of an input picture as a natural language sentence when given an image as input rather than a phrase. This study employs the following formula to increase the likelihood of a proper picture description being provided:

required for computer programming. Even though they are merely the groundwork steps before you start writing, they are crucial. Whether you're trying to learn about neural networks

In this formula, I is the input picture, is a model parameter, and S is the image's precise "translation" representation. Since S represents the whole list of possible results from a statement, its size is unbounded. In this study, we model all the probabilities, from S0 to SN, using the chain rule. When N is this particular integer, for instance:

$$i_t = \sigma\left(W_{ix}x_t + W_{im}m_{t-1}\right) f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1}\right) o_t = \sigma\left(W_{ox}x_t + W_{om}m_{t-1}\right)$$

$$\log p\left(S \mid I\right) = \sum \log p\left(S_t \mid I, S_0, \ldots, S_{t-1}\right)$$

is a set used for training (see Section 4 for more information on training) and we utilize stochastic gradient descent to maximize the set's cumulative probability.

With respect to the combined probability p(St|I, S0. After a fresh input xt is seen, the information is solved using the following nonlinear function f:

A parameter in training is a W matrix. This item

These AND gates, when given a value larger than 1, may effectively prevent the gradient explosion or disappearance induced by weights greater than 1, allowing LSTM to engage in high-intensity, large-scale training. The last equation requires an input value, Mt, which is the output value of the function Softmax.

f ht ht, ht 1 ht, xt

This study presents two major selection problems—the particular form of f and the means by which images and natural language phrases may be input and output—to provide a clearer illustration of the RNN's underlying working concept. The research used a Long Short-Term Memory (LSTM) network for f, which, by the standards of the time, performs exceptionally well on sequence problems like translation. In the next part, this article will elaborate on this model's (LSTM) specifics.

This research achieves the goal of representing pictures by using Convolutional Neural Networks (CNN). In the realm of target identification and detection, CNN is the cutting edge technique that has recently seen widespread usage in academic research. These paper-selected CNNs use a novel batch normalizing technique. They were also demonstrated to be transferable to other tasks, such as scene categorization, through transfer learning. An embedding model serves as a representation for these terms.

Sentence Composition System Based on LSTM

The following formula (3-3) illustrates that the particular form of f is determined by the network's capacity to handle image input and gradient descent, which is the primary challenge for all users of CNNs and RNNs. LSTM, a recurrent neural network with a

ISSN 2347–3657
Volume 6, Issue 2, 2018

novel data structure, was developed to address this issue and has since been put to use in a wide variety of translation and sequence creation applications.

The LSTM model's storage unit c is its most original feature since it can store the output sequence for each input (see Figure 24). The "gate" gives the order to carry it out. If the gate's value is 1, the gate control layer will return a 1. When the gate is set to 0, this value does not change. The following three gates regulate c's overall behavior: First, the Forget gate f determines if the contents of the cell are inherited; second, the Input gate i determines whether the input information is read; and third, the produced gate o determines whether a new value is produced. Following is a definition [17] of gate and cell input and output:

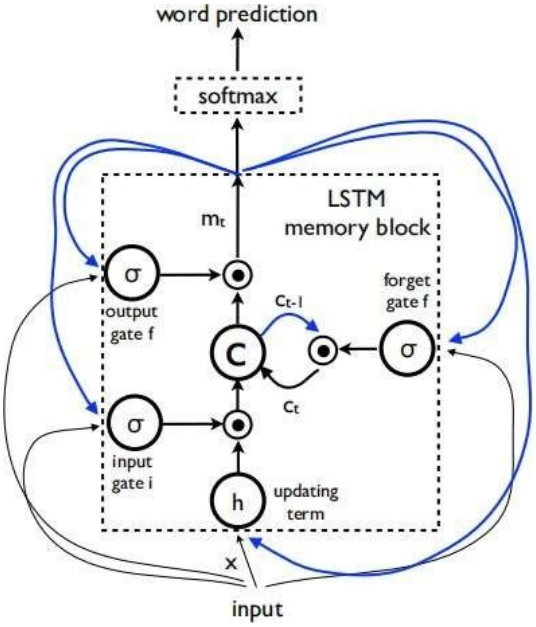


Figure 24.
LSTM model diagram

As may be seen in FIG. There is a storage unit c and three control gates in the LSTM memory block. The blue line represents the feedback loop: the output m at time t-1 is received by three gates and fed back to the memory c at time t, and the output value mt at time t will be fed back to the Softmax function alongside the output m at time t from the memory..

Add Attention Mechanism

The study of human eyesight was the inspiration for Attention Mechanism. Humans, according to cognitive science, only pay attention to a subset of all observable information because of processing
Among them, the length of Source is denoted by the formula Lx=||Source||, the significance of which was previously explained. It may be difficult to see this structure that reflects the essential idea in the machine translation example given above because, in the process of calculating Attention, the Key and Value in Source are combined into one, pointing to the same thing: the semantic code corresponding to each word in the input sentence (RNN refers to the bottlenecks. A common name for the system described above is the "attention mechanism" [20]. When reading, for instance, most individuals only focus on a fraction of the words on the page [18]. In conclusion, the attention mechanism is responsible for prioritizing where limited processing resources should be used and selecting which component of the input requires attention.

This is one method to analyze the Attention process (see Figure 25): Let's assume the Source data is organized as a set of Key, Value> pairs. In the present moment, given a target element Query, by determining Query The final Attention value is calculated by first determining the weight coefficient of each Key corresponding to the Value, then calculating the total of these weights. Therefore, the Attention mechanism is a weighted sum of the Value values of the items in the Source, and the weight coefficient of the corresponding Value is determined by the Query and the Key [18]. In other words, the formula for its core concept is as follows:

Take note, "Query, Sorce"

address. The content will be extracted, with the significance of the extraction evaluated by the similarity between Query and Key; the Value will then be weighted and added in order to extract the final Value value, which is the Attention value [20].

If most of the existing approaches are abstracted, the Attention mechanism's precise calculation can be boiled down to two steps: (1) determining the weight coefficient based on the Query and Key, and (2) determining the Value based on the weight coefficient. Summation using weights [21]. In the first step, similarity or correlation is computed between the two inputs (the Query and the Key), and in the second step, the raw score from the first step is normalized. This allows us to conceptually separate the Attention calculation process into

$$Lx = \sum Similarity\ (Query,\ Key_i)*Value_{ii=1}$$

three stages [21] as shown in Figure 26 below.

hidden layer state). From a conceptual standpoint, however, the notion that Attention is the process of identifying and concentrating on a subset of relevant information while disregarding the vast majority of irrelevant data remains unchanged [19]. The act of concentration is mirrored in the determination of the significance factor. The larger the weight, the more attention is paid to the Value value that it represents; in other words, the weight indicates the significance

of the information, and the Value is what that information is.

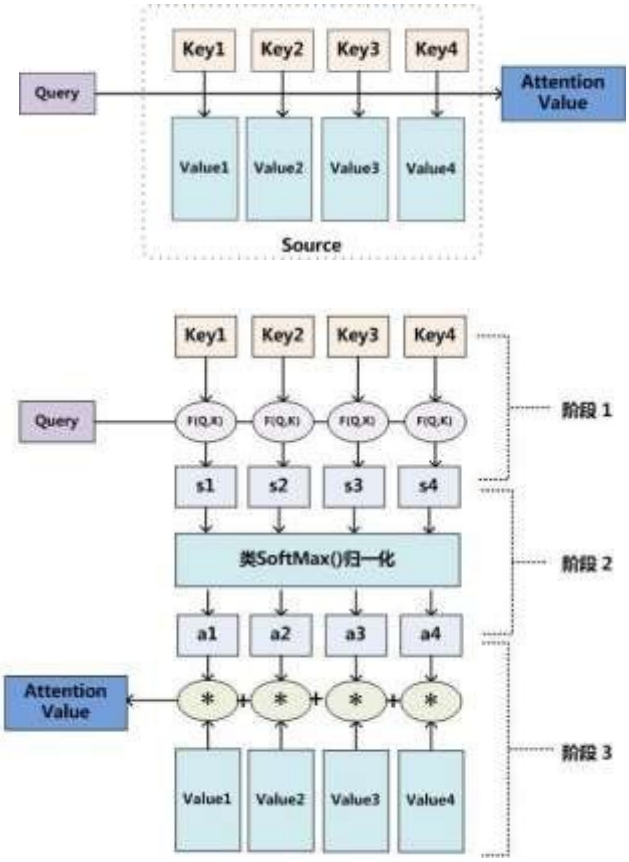


Figure 26. Schematic diagram of the three-stage calculation process ofthe Attention mechanism

what kind of data is being saved. When addressing, we check how closely Query matches the address of the element Key in memory. Soft addressing is different from traditional addressing in that it just looks for data inside the storage itself, rather than across several keys.

In the first stage, different functions and calculation mechanisms can be introduced, and the similarity or correlation between the two can be calculated according to Query and a certain Key_i. The most common methods include: calculating the vector dot product of the two, calculating the The vector Cosine similarity is either evaluated by introducing an additional neural network [18], that is, the following formula:

$$Similarity(Query, Key_i) = MLP(Query \cdot Key_i) \quad (3\text{-}13)$$

The score generated in the first stage varies accordingto the specific method of generation, and its value range is also different. In the second stage, a calculation method similar to SoftMax is introduced to convert the scores of the first stage numerically. On the one hand, it can be normalized, and the original calculated scores can be sorted into a probability distribution where the sum of the weights of all elements is 1; on the other hand, it can also be the weights of important elements are more highlighted

by the intrinsic mechanism of SoftMax [18]. That is, it isgenerally calculated by the following formula:

$$Similarity(Query, Key_i) = MLP(Query \cdot Key_i) \quad (3\text{-}14)$$

The calculation result of the second stage a_i is the corresponding weight coefficient, and then the weighted summation can be used to obtain the Attention value :

$$Attention(Query, Sorce) = \sum_{j=1}^{L_x} a_i \cdot Value_i$$

Through the above three-stage calculation, the Attention value for Query can be obtained. At present, most of the specific attention mechanism calculation methods conform to the above-mentioned three-stage abstract calculation process.

Training Process

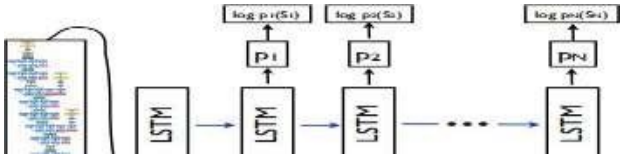
By training an LSTM model to predict each word in thesentence after seeing the image, and p(St|I, S0,…). Inorder to achieve these functions, we need to expand the LSTM - a LSTM memory for image input, and

$$x_{-1} = CNN(I)$$

features, and the features are stored in the convolution kernel. Then, each feature is compared through a weighted calculation, and the feature vector with high comprehensive weight is selected for training. Find the corresponding labels in a centralized manner, and enter the LSTM for word sorting training. During the training process, the parameter values of the convolution kernels of the CNN are continuously adjusted through logistic

ofthe output natural language sentences. In this way, a word input for each sentence, so that all LSTM models can share the same parameters and outputs. All periodic connections are transformed into unrolled versions of feedforward connections. In another easy-to-understand way, we use I to represent the input image, and a vector groupS = (S0,...,SN) to represent each word of this image, and the expansion process is [17]:

system model with the function of seeing pictures and talking through training is obtained, which is the result of the design of this project.



International Journal of Information Technology & Computer Engineering

ISN NO: 2347 – 3657
Volume 6, Issue 2, 2018

$$p_{t+1} = LSTM\left(x_t\right), t \in \{0...N-1\}$$

Figure 27. LSTM work timeline expanded view

4. Training and Testing ExperimentalResults

Each feature vector St represents a word, and the dimension of each vector is equal to the size of the "thesaurus". It is worth noting that the first value S0 represents a special start word, while Sn represents a special stop word, which respectively represent the beginning and end of the sentence. It is worth noting

$$L\left(I,S\right) = -\sum_{t=1} \log p_t\left(S_t\right),$$

3.5. Chapter Summary

Section 4.3. An extensive evaluation protocol, and the

This chapter mainly describes the specific constructionof the overall model. First, CNN is used to extract image

because BLEU always prefers 3. A more detailed discussion of metrics can be found in, and research groups studying this topic have reported other metrics deemed more suitable for evaluating titles.

In conclusion, the advantage of using proxies to describe is that known ranking metrics can be used, which is more convenient. On the other hand, converting a description generation task into a ranking task is not ideal: as the complexity of the image description increases, along with its dictionary, the number of possible sentencesgrows exponentially with the size of the dictionary, suitable for the prediction of new images. The number of defining sentences also goes down, and the potential computational complexity for efficient evaluation of the large number of sentences stored per image. The same is true in language recognition. Whether the sentence corresponding to a given acoustic sequence can be generated is the first problem to be solved. State-of-the-

that, by sending out the terminator Sn, the LSTM model receives the signal that the sentence has been constructed, thereby mapping both the image and the word into the same space. The image recognition adopts theconvolutional neural network CNN, and the label word is embedded in We. The image I is input once and only once at t = −1, giving the LSTM model the input image content. After testing and experiments, inputting additional image content at each time step will lead to large errors, because the network will easily identify too many "details" in the image, resulting in overfitting. where the resulting loss is the sum of the negative log-likelihoods of the correct wordat each time step [17], as follows:

$$N$$

4.1. Evaluation Indicators

Although it is sometimes unclear whether a description should be considered successful, several evaluation metrics have been proposed in the prior art. Obviously the most accurate method is to ask the rater to directly rate each output title for a given image, i.e. ask the rater to rate each generated sentence on a scale from 1 to 41. For this metric, this paper sets up an "Amazon Mechanical Turk" experiment. Each picture is graded by 2 students. The consensus level between the two is usually 65%. If there isdisagreement, the scores are simply averaged and recorded as a score. For ANOVA, perform bootstrapping (replace and calculate mean/standard deviation for resampling results). Like it, the scores reported in this paper are greater than or equal to a set of predefined thresholds.

Although this metric has some obvious shortcomings, it has been shown to correlate well with human evaluations. In this work, this is also confirmed, as shown in

output produced by our system, can be found at http://nic.droppages.com/.

Based on the objective function in (1), complex models can be transcribed without BLEU. The choice about

model selection and hyperparameter tuning is performedby the value of perplexity, but this paper does not report it,

artmethods for this task are now generative, producing sentences from a large dictionary.

Now that our model can generate descriptions of reasonable quality, and despite the ambiguity in evaluating image descriptions (where there may be multiple invalid descriptions), we believe that this paper should focus on generating evaluation metrics for the task rather than ranking.

Selection and Testing of Datasets

For evaluation, this paper uses a large dataset consisting of images and English sentences describing those images. The data set statistics are as follows:

Table 1. Statistics of major datasets

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 6, Issue 2, 2018

Dataset name	Size		
	Train	Valid	Test
Pascal VOC 2008	-	-	1000
Flickr8k	6000	1000	1000
Flickr30k	28000	1000	1000
MSCOCO	82783	40504	40775
SBU	1M	-	-

With the exception of SBU, the sentences for each image are relatively visual and unbiased. SBU is the composition of the description given when the image is uploaded to Flickr. Therefore, the requirements of this performance cannot be guaranteed, and it can be seen that the data set has some interference. The Pascal dataset is traditionally only used for testing after the system hasbeen trained on different data (such as any of the other four datasets), so it is not suitable as the first dataset for training. So we drop SBU, use the other 1000 images for testing, and use the rest for training. Use it to report resultsin the next section.

Results

Because the model in this paper is a data-driven, well- trained end-to-end system and is trained on a rich dataset, this paper hopes to answer questions such as "how doesthe size of the dataset affect generalization", "what kind of Transfer learning makes it happen" and "How to deal with weakly labeled examples". Therefore, we conduct experiments on 5 different datasets, which enable us to gain a deeper understanding of our model, as detailed in Section 3.2.

Training Process and Its Details

Many of the problems we encountered while training our models in this article were related to overfitting. The relatively large dataset of ImageNet is used for data-driven to solve the problem of difficult assignment and description.

Therefore, even with reasonably good results, the advantage of this method over the currently widely used engineering methods will only increase over the next few years as the size of the training set increases.

Nonetheless, this article considers several ways to deal with overfitting. The most obvious way to avoid overfitting is to initialize the weights of the CNN component of this system to a pre-trained model (e.g., on ImageNet). This paper does this in all experiments, and it does help a lot in generalization. Another set of weights that can be reasonably initialized is We, the word embedding. This paper attempts to initialize them from a large news corpus, but no significant effect is observed, and it is decided not to initialize them for simplicity. Finally, the paper employs some techniques to avoid model-level overfitting. Dropout and the overall modelgave some improvements in BLEU scores, which are mentioned throughout the paper.

We train all weight sets using fixed learning rate and momentum-free stochastic gradient descent. All weights are initialized randomly except the weights of CNN, which are not changed in this paper because changing them would have more negative effects. This paper uses 512-dimensional embedding and LSTM memory size.

In addition, we perform basic tokenization preprocessing on the description, keeping all words that appear at least 5 times in the training set.

Generate Results

This paper reports the main results on all relevant datasets in Table 1 and Table 2 below. The state-of-the-art results of PASCAL and SBU do not use deep learning- based image features, so it can be said that this change alone can improve these scores by a large amount. The Flickr dataset has been used recently, but mostly for evaluation in retrieval frameworks. A notable exception is that they are retrieved and generated simultaneously and yield the best performance on the current Flickr dataset.

The human scores in Table 2 are calculated by comparingone of the "Human" titles with the other 4 "Human" titles. This paper does this for each of these five raters and averages their BLEU scores. Given that the BLEU score is calculated from 5 reference sentences instead of 4, this paper adds back the average difference of 5 referencesentences instead of 4 reference sentences to the score.

Given the significant progress the field has made over the past few years, this paper argues that it makes more sense to report on BLEU-4, the standard for moving forward in machine translation. Despite my constant efforts to get better test results, the model in this paper outperforms human raters. However, when human raters are used to evaluate the captions (i.e. when classmates are asked to give captions to the images), our model performs worse, suggesting that more work is needed to obtain better metrics. On the official test set (labels are only available through the official website), our model has a BLEU-4 score of 27.2.

Table 2. Test score results on MSCOCO dataset

Standard	BLEU-4	METEOR	CIDER
NIC	27.7	23.7	85.5
Random	4.6	9.0	5.1
Nearset Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Result Analysis

1) After testing, the results will be sampled and illustrated in this article. First of all, in addition to the BLEU score, this paper conducts a manual

ISSN 2347–3657
Volume 6, Issue 2, 2018

comparative analysis of the results and standards. The conclusion is that the system performs extremely well in terms of grammar. Only 24 pictures out of 1000 test pictures have basic grammar errors. For example, tense grammatical errors, etc. This performance is actually expected in this article, indicating that the design of the LSTM model in this article is as perfect as this article imagines. However, there are a lot of errors in the description of specific characteristics of things. In addition, some errors will also occur when processing some special images. The following lists several cases with errors or unrecognizable:

2) The tense grammar is wrong

Figure 28. Common test chart

As shown in the figure, this is an image of a football match, specifically Neymar in yellow jersey and two players in white jersey grabbing the ball. Due to the limitations of this model, only the most basic and most important ones will be generated. Description, that is, the content is: Three guys are playing football. Here is the result given by the program system:

Figure 29. Schematic diagram of test results

Obviously not similar to the standard answer, but this will not affect the score too much, because there are many synonyms in the library. The bigger problem is the grammar, obviously there is a missing "is" here, which constitutes a grammar error, which probably won't produce a comprehension error in this picture, but will be greatly discounted in the scoring. Whether it's changed to "is playing" or just a simple present tense of "play" is correct in the evaluation. According to the analysis of the code and process, the possible problem is that there is an error in the feature vector extraction process of CNN, resulting in the use of "playing" during the process, and it is not considered to add "is" before. This is unnecessary when extracting the tense of the simple present tense. Perhaps a requirement should be set in advance that when the "-ing" tense occurs, the verb "be" is added before the word.

3) Anime image recognition error

Figure 30. Schematic diagram of test animation picture results

To evaluate MSCOCO's capacity to detect animation pictures, this article uses a picture that is not included in the MSCOCO atlas but was instead randomly obtained from the Internet.

There is obviously a major mistake in the attribution. Mistaking a cat for a man or a mouse for a banana is an extreme departure from the intended meaning of the image. This system's present identification capacity for animation images has been thoroughly tested and is accurate. There are significant problems, and no workable solutions have been proposed as of yet. Due to the fact that the data set employed for the purposes of this work consists entirely of genuine photographs, the system cannot be trained to recognize animated images. It has also been shown that training with

ISSN 2347–3657
Volume 6, Issue 2, 2018

the inclusion of anime visuals is impossible. This is due to the fact that the "cat" picture has already been programmed with the correct parameters according to the Luo Ji regression algorithm. Since the animation picture and the actual image vary too much, adding more animation images to the training set would simply increase the identification error.

It also considerably increases the complexity of training as a result of the more complicated aspects such as animation style in animation images. There is currently no workable solution shown here. Future, in-depth study is needed to tackle this problem. Now that this content has been edited and proofread, the continuous tense voice may be presented accurately for the first time.

4) **Recognizing monochrome images**

Black-and-white picture identification is still rather accurate, as can be shown, however there may be some issues with recognizing relatively tiny things. "Cake" may really be "tea" or "foods" in this image. This, it seems, is because chromatic aberration in black and white images is less noticeable.

not as blatant as in color RPG images. In light of the pixel issue, this paper suggests that mistakes might occur even if artificial recognition is used. Allow this article to serve as an example and label it. Such clarity makes it difficult for this article to tell whether the liquid contained inside this cup is wine or tea, necessitating more investigation. As may be seen, the system described in this study is a long way from reaching that pinnacle.

In spite of subsequent adjustments to the tense grammar, this article continues to consider the case to have been a success.

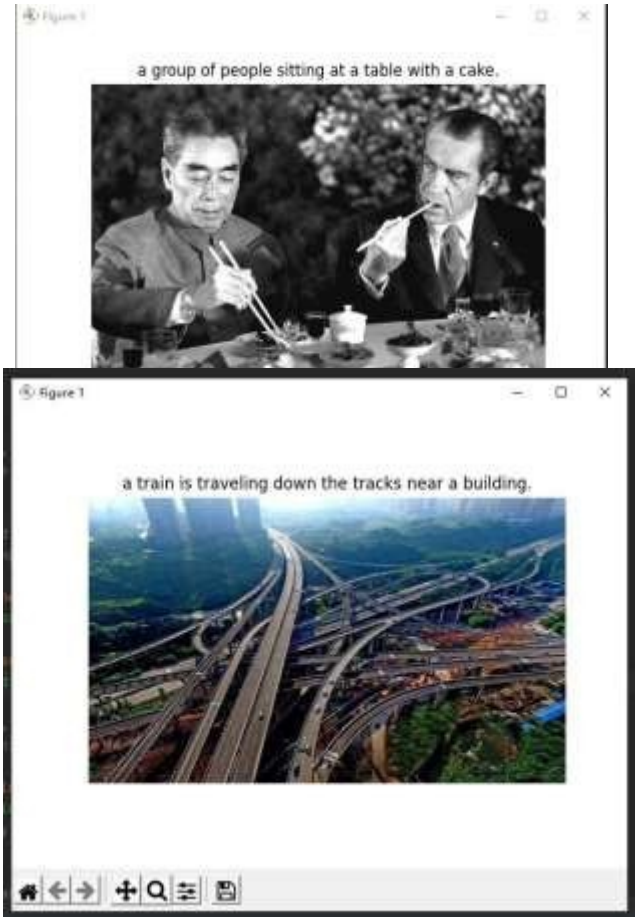


Figure 31. Schematic diagram of the test results of black and white pictures

4) Complex images

Figure 32. Schematic diagram of the result of testing complex pictures

Images with too many components or too much material are still challenging for the algorithm to detect. Even though the MSCOCO data set, which has rather rich content, has been utilized for training, it is clear that the discriminative capacity is not sufficient to distinguish such complicated pictures. Training on larger and more complicated datasets, however, is expected to provide bigger progress in this area.

This page can't be updated quickly for technical and network-related reasons, however the situation is not hopeless. This piece is based on faith.

Summary of This Chapter

After reading this article, you'll know which kind of photos are most likely to cause problems for the system. The majority of the photographs are also identified without serious issues, which is consistent with the aims of this essay. This work has improved and dealt with several issues that it can enhance or fix quickly (such as the capacity to describe temporal grammar). The current performance and indicators have fully reached the expected value of this article, and have also achieved The basic purpose of this subject. However, some other problems are still difficult to solve at the present time and may need to be solved in this article after extensive study and research, and under better equipment conditions.

References

This is [1] Liu Siyang. Studying the Effects of Combining Syntactic and Semantic Information in Sequence-Tree Encoders for Natural Language Reasoning [D]. 2018 Zhejiang University.

[2] Guo Jimin. Study and use of deep neural network-based techniques for object recognition [D]. Electronic Science & Technology University of China (2018).

In reference to Zhang Yanqi (3rd). Imagine a deep learning-based Chinese semantic understanding [D]. Technology in Harbin, China, 2017.

Xidian University, 2019. [4] Ma Xinrui. Primitive Feature Analysis in Image Recognition Algorithm Research [D].

Song Jiantao, Reference No. 5. [D] Develop a system for early warning based on a platform for tracking the provenance of agricultural products. Technology at Beijing University, 2018.

Han Guo Feng, reference number 6. Transfer learning-based personalized recommendation system for tourism attractions [D]. Science & Technology at Shaanxi Province, 2019.

7] Guo Fei. Target identification studies using deep learning for mechanical components [D]. Technology at Lanzhou University, 2019.

As in [8] Chen Jiaming. Technology for correcting errors in satellite location via the use of a convolutional neural network: [D] research and simulation. Published in 2018 by the Beijing University of Posts and Telecommunications.

Tensorflow-based Recurrent Neural Network Model for Predicting Shanghai's Air Quality [D], Liu Lei et al., Shanghai Normal University, 2019.

Gao Maoting and Xu Binyuan [10]. The use of a recurrent neural network in a recommendation system [J]. 2019;45(8):198-202+209 Computer Engineering.

A Generative Adversarial Network-Based Algorithm for Generating Images with Textual Descriptions, by Wu Haoyu [D], Nanjing Normal University, 2019.

Referring to Fu Yuan, [12]. Information & Communication, 2019(08): 137-138, presents "A method and system for testing RDMA data

ISSN 2347–3657
Volume 6, Issue 2, 2018

transmission on Tensorflow software."
[13] Yin Yuecheng. Turning experiments using DT4E pure iron [D] materials. Technology at Dalian, 2019.
Specifically, Lv Ruru (Ref. 14). Studying the digital copier's original data collecting and processing technology [D]. 2011; Nanjing; Nanjing Forestry University.
As cited by Xie Pengfei in [15]. Deep learning (D)-based geostatistical inversion technique. University of Yangtze, Year 2019.
This is Zhi Shuaifeng, reference number 16. The use of convolutional neural networks in 3D object recognition research [D]. Technology in Defense, National University of, 2017.
Dou Min [17]. CNN and LSTM-based video semantic analysis system [D] design and implementation. The School of Journalism and Mass Communication at Nanjing University
Y. Xiao, H. Jixiang, Y. Zheng, B. Wang. The use of several attention processes and external information to generate a picture description [P]. Dated May11, 2021 in Liaoning Province, CN112784848A.
Source: [19] Ge Hongwei and Yan Zehang. [P] is a technique for automatically creating picture captions using multimodal attention. Province of Liaoning, CN108829677B, May 7, 2021.
Hu Fei, Peng Liang, Zhong Wei, Fang Li, Ye Long, Zhang Qin, and 20 others. Beijing: CN112733944A, 2021-04-30. Method, apparatus, and media for object recognition based on picture and category attention [P].
Legal Question Answering System Research (D) by Zhou Yiwen, Hunan University, 201.