Image-Based Storytelling for Tourist Using Deep Learning

Yulin Zhu

A project report submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2021

School of Engineering, Computer & Mathematical Sciences

Abstract

In order to describe a journey, it can be generated a story from a series of digital photograhs. Most of the existing methods focus on descriptions of the specific content of a single image, such as image captioning, which lack of the correlation between the images and the relevance of spatial dimensions. To this end, in this report, we propose a novel visual storytelling architecture based on computer vision in terms of object detection from photos. By extracting the information of special objects from the images, combining the changes in spatiotemporal domain, and filling in the predetermined template, we generate a travel diary. The robustness of the algorithm is improved by optimizing the data set acquired by amplification.

In this project, compared with traditional image captioning, our contribution is to effectively connect the correlation between images and the potential background meanings of the story. The contributions of this report are: (1) Innovative use of preset templates to generate travel diaries from image streams, (2) introduce context to the picture stream and maintain the story background in a long sequence of events, (3) adjust the gray value of the picture to expand the data, (4) the image preprocessing method shortens the expected training time.

Keywords: Storytelling, object detection, CNN, data enhancement

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Yulin Zhu Date: 29 Oct 2021

Acknowledgment

l would like to express my gratitude to all those who helped me during the writing of this report. First of all, I would like to gratitude my beloved parents for their emotion support and great confidence. Owing to the unselfish and generous sponsor from them, I have this invaluable opportunity to complete my masters degree in the Auckland University of Technology (AUT), New Zealand.

Secondly, l also owns a special debt of gratitude to my primary supervisor Dr. Wei Qi Yan for his constant encouragement and guidance. In this study, he provided me with professional knowledge support and careful guidance, without his consistent and illuminating instruction， this report could not have reached its present form. Finally, l am also deeply indebted to all the other teachers and friends for their direct and indirect help to me.

Yulin Zhu

Auckland, New Zealand

October 2021

Chapter 1
Introduction

This chapter is composed of five parts: The first part introduces the background and motivations, the second part includes is the objectives, followed by the research question, contributions, and structure of this report.

1.1 Background and Motivation

With the rapid development of Internet, a large amount of multimedia data is accumulated. One of the most difficult areas is how to let the computer understand the information contained in a picture. A picture from a computer is a large amount of color and optical information into a huge matrix of numbers, known as pixels. With the development of current technology, we have taught computers to look and discrible a image, known as image caption. Therefore, the machine could describe the surface information contained in an image, just like to discribe “a dog is running”. But when comparing what storytelling is different from image caption, we could say that taking a picture is only record that what did we see, but the storytelling is to record how we feel and what we thought. Which could give us more option to record and share our memory.

For those who love to travel, they feel and discover many things that have never been seen before, such as animals that have only appeared on TV, delicious food described in books, and places of historic interest that have gone through thousands of years. These things are not only an object or a building in the picture, but also the background story that every visitor wants to know and the same or different feelings after experiencing them. The image recorded by a photograph often has many special meanings (Smilevski et al, 2018).

Therefore, how to record and convey the story and emotion behind a photo has become one of the new problems worth exploring. Travel photos contain much more information than just looking at image and speaking. The image-based storytelling is a kind of artificial intelligence, which developed from the deep learning system. The image-based storytelling methods take use of the picture(s) to generate a summary or conclusion based on the image analysis result. By finding key object information in photos and combining with big data, we can better record the background stories of related object, so as to enrich the words describing photos. The image-based storytelling involves both the analysis of the image and the generation of human natural language. People will take the time to document a trip because the words behind the pictures is better to help them to describe, record and share a unique memory. Taking object detection as a starting point

and integrating the massive object description template brought to us by big data, we use photos to generate travel diaries more conveniently and quickly. We hope that this method will help people with disabilities as well. This study will demonstrate using deep learning to automatically identify incoming images extract the key features in the photos, and generate travel diaries according to the timeline.

1.2 Research Questions

According to some demonstrations of image caption, we find that at present, image captioning tends to stay in the description of the content of a single picture, it is difficult to detect a large number of emotions, experiences, story background and other information contained behind the picture. Therefore, based on the above problems, the following issues will be studied and discussed in this report:

(1) *How to use deep learning technology to quickly extract key objects from photos?*

(2) *How to augment training data when the number of photo samples is small?*

(3) *What impact does Transformer and Attention Mode have on the original CNN detection algorithm?*

(4) *How to extract emotion information from existing text using NLP?*

(5) *Explore the benefits this technology can bring to people with disabilities*

The main purpose of this study is to extract the key targets in the photos by using the object detection method of deep learning and analyze the travel information of one or more photos by combining with big data, so as to generate travel stories containing emotions.

1.3 Contributions

The focus of this study is to use object detection in deep learning training to detect special objects from the given photos, and obtain story descriptions of one or more photos

containing emotions by using template-based languages. The target detection method will also be compared with Transformer. Throughout this report, we contribute to: (1) Visual object detection using CNN and Transformer, (2) analyzing object detection algorithms suitable for specific data sets, (3) analyzing the emotion contained in the event using the POS tagging method based on HMM, (4) data augmentation to a specific data set, (5) generating travel stories based on provided object, timelines, and emotional information, (6) the possibility that this technology can bring to people with disabilities.

1.4 Objectives of This Report

Our first goal is to train and use deep learning based on CNN for object detection, during which we will evaluate the possible improvements and comparisons that Transformer can bring to object detection. NLP is used for emotional analysis of the story, so that the story can maintain a stable and orderly emotional background according to the development of the timeline. Finally, using the objects in the photos, based on the upper bound text of the timeline, the prefabricated templates are used to generate the image-based travel stories. In addition, we hope to use mature text-to-speech methods to help disabled people conveniently.

1.5 Structure of This Report

The structure of this report is described as follows:

- In Chapter 2, the focus will be on the research and application of CNN and Transformer in object detection. We will also examine the basic implementation of image caption, as well as discuss the advantages and shortages of different text generation methods.
- In Chapter 3, we will introduce the research methods. Experimental design and resultant comparisons will be present in this chapter.
- In Chapter 4, we will implement the proposed algorithms, collect experimental data and demonstrate the research outcomes in the form of figures and tables. Additionally, the limitations of these proposed methods will be detailed.

- In Chapter 5, we will summarize and analyze the experimental results.
- We will draw the conclusion and state our future work in Chapter 6.

Chapter 2
Literature Review

The focus of this report is on visual object detection and generation of story by using deep learning, this chapter will introduce a plenty of traditional methods and the relevant knowledge of deep learning.

2.1 Introduction

Pertaining to the tourists who like travels, they enjoy the beautiful scenery, broaden their horizons and explore the unknown world. It's also important to record your journey and find emotional resonance. In the same place, people will have similar feelings and emotions. Sharing and recording these feelings has become a part of travel.

The moment the photos were taken to record the specific time, scenery, people and things, which may be the beautiful scenery, but also the feelings brought to people by the scene at this moment (Smilevski et al., 2018). What is recorded behind a photo can be written down as words to describe a scene or a story. By combining storytelling, we make ourselves and others better understand a culture, convey emotions and make a sightseeing trip more meaningful.

2.2 Deep Learning

Before starting studying neural networks, as a summary, we will discuss the relationship between artificial intelligence, machine learning and deep learning.

In the summer of 1956, an American conference at Dartmouth University, chaired by John McCarthy, proposed the concept of artificial intelligence, using the then newly emerging computers to build complex machines with essential properties similar to human intelligence. Many years later, it was identified as the starting point of global artificial intelligence research (Copeland, 2015). Artificial intelligence is a quest to simulate human behavior, thinking and action. The foundation of artificial intelligence is based on philosophy, mathematics, linguistics, psychology, and computer engineering. The development process of artificial intelligence has also experienced the discovery, exploration, development, in-depth applications, and others (Russell & Norvig, 2021).

As the expectation of computer science is become higher, the problems to solve are more and more complex. Machine learning utilities algorithms to analyze data, learn from it, and then use that data to resolve predictions about specified issue (Neapolitan & Jiang, 2018). Computing algorithms are trained with large amounts of data to perform a specific

task. Machine learning algorithms allow computers to perform tasks without a lot of programming. Machine learning is also a subset of artificial intelligence (Sharp, Ak & Hedberg, 2018). The widely cited definition of machine learning gives the best explanation. (Mitchell, 1997).

Machine learning is a subset of artificial intelligence (AI) that enables machines to automatically learn the benefits of concepts and knowledge without explicit programming. Deep learning is based on a collection of machine learning algorithms, such as linear regression, logistic regression, and multilayer perceptrons. In the training stage, a large amount of data will be used to extract features with appropriate mathematical formulas; in the reasoning stage, the key features previously extracted from a large amount of data will be used to fit and mark new undisclosed data (Dargan et al, 2020). The method of learning can be supervised (Caruana & Niculescu-Mizil, 2006) or unsupervised (Hinton & Sejnowski, 1999; Hinton, 2012).

Deep learning is an approach to mimic the way the human brain works with neurons, using nested layers of concepts to express and achieve great power and flexibility. Artificial neural networks (ANNs) consist of a hierarchical structure of algorithms. Each feature is represented by a node (Goodfellow et al., 2016).

Deep learning is a mathematical problem. Machine learning is the methodology of deep learning, mathematics is the theoretical support behind it. The goal of machine learning is to find a mathematical function that can fit the relationship between input data and output as accurately as possible. This principle is called universal approximation theorem. The principle is simple -- a neural network can fit any function, no matter how complex its expression is (Nielsen, 2016). Cybenko proved that multilayer perceptrons such as feed-forward neural networks with at least one hidden layer can approximate any continuous function (Cybenko, 1989). The universality property has also been proved in the case of convolutional neural networks that shows that convolutional neural networks can approximate any continuous function to an arbitrary accuracy when the depth of the neural network is large enough (Zhou, 2020).

The development platforms for deep learning include Python (Ketkar & Santana, 2017; Chollet, 2017), TensorFlow (Raschka & Mirjalili, 2017), and MATLAB (Kim,

2017; Vedaldi & Lenc, 2015). The software has plentiful open-source library support, good community discussion, and easy-to-learn features that make them more friendly to researchers.

At present, deep learning has been widely applied. Artificial intelligence is based on algorithms and learns from a large number of characteristic data. Based on the approximation of everything theorem, the algorithm should be able to find the matching rule exactly as long as it has enough data. For example, in the medical area, Zhang et al., has taken use of deep learning in modern drug discovery under big data era, and clinical drug candidates discovered by computational methods are also offered (Zhang et al., 2017). There are also applications in medical diagnostics, using deep learning DenseNet and VGG 19 to analyze X-ray images to aid in COVID-19 diagnosis in the context of the global COVID-19 pandemic (Oh, Park & Ye, 2020; Hemdan, Shouman & Karar, 2020). More importantly, deep learning has also been applied to image classification, facial recognition technology and MNIST for handwritten number recognition (Chan et al., 2015), as well as data enhancement when processing training data (Perez & Wang, 2017).

2.3 Convolution Neural Network

As the foundation of deep learning, artificial neural networks are largely inspired by the way biological nervous systems, such as the human brain, work. It consists of a large number of computing nodes that work like neurons, receiving input and output results at different thresholds. For the traditional artificial neural network, the representative one is fully connected neural network. LeCun et al. introduced the basic convolutional neural network layer, which includes the convolutional layer, pooling layer and full connection layer (LeCun et al., 1998).

Usually, the image is treated as input, and usually the two-dimensional matrix is converted directly into a one-dimensional vector (O'Shea & Nash, 2015). Compared with the traditional fully connected neural network, convolutional neural network has more advantages in image processing. Therefore, CNN has also become a reliable method in face detection, picture recognition and video recognition (Khan et al., 2017; Albawi et al.,

2017). It mainly solves the problems that spatial information will be lost in the process of expanding the image into a vector, and the direct expansion into a vector will lead to too many parameters, slow training and easy over-fitting of the network. (Albawi et al., 2017). In particular, the study of Huang et al. proved that DenseNets greatly reduced the number of parameters (Huang et al., 2017).

CNN model based on MNIST and CIFAR-10 datasets is evaluated with image recognition and datasets. Specifically, the operation and sliding of convolution kernel are used to extract feature images. A feature map is usually a local receptive field for a region of the image.

$$N \times N * f \times f = N - F + 1 \tag{2.1}$$

where $N \times N$ images are convolved with $f \times f$ filters to extract image features. The study also uses ReLU as the activation function and Dropout to reduce overfitting of the model (Chauhan et al., 2018). Guo et al. 's study also proved that using CNN for image classification is a common and relatively simple method, and the influence of different learning rates and solving strategies on model training was also studied based on MNIST and CIFAR-10 datasets (Guo et al., 2017).

The influence of full connection layer is based on the image classification results of convolutional neural networks. The results include the shallow CNN required fewer nodes than deep CNN. Shallow CNN is suitable for wider datasets, while deep CNN is more suitable over deeper datasets (Basha et al., 2020).

2.4 Recurrent Neural Network

CNN (Convolution Neural Network) and RNN (Recurrent Neural Network) are the two discriminative structures in deep learning (Schmidhuber, 2015). Recurrent neural networks are a class of networks that take sequential data as input, i.e., the previous input is related to the later input, and are commonly used for natural language related problems. RNN could be used in parsing and language modeling, and had good parsing ability both in English and Chinese (Dyer et al., 2016). A simple recurrent neural network, for example, consists of an input layer, a hidden layer and an output layer (Goodfellow et al.,

2016). Among them, Bidirectional RNN (Bi-RNN) (Schuster & Paliwal, 1997; Dhyani & Kumar, 2021) and Long Short-Term Memory networks (LSTM) (Greff et al., 2016; Sundermeyer et al., 2012) are the common recurrent neural networks.

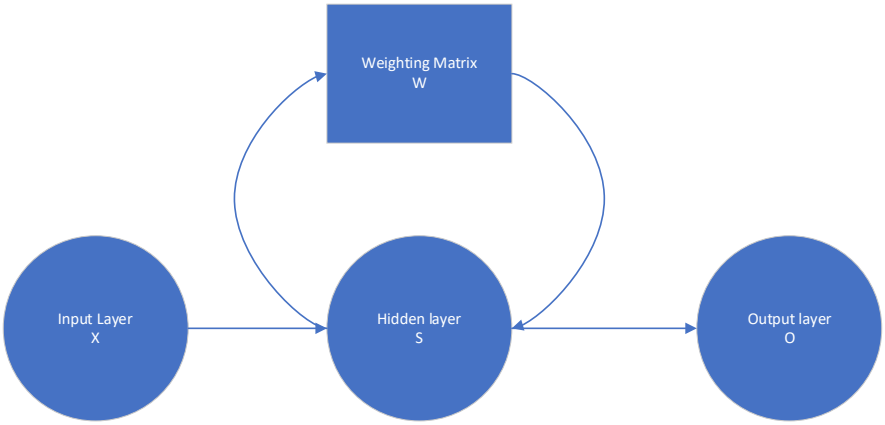

Fig 2.1: The RNN model

RNN was designed for text classification. It is noted that RNN-based text classification has excellent performance on document-level datasets, outperformed than the state-of-the-art methods. The use of RNN allows a good connection between a word and the meaning of its left and right sides compared to traditional neural models (Lai, 2015).

The Long short-term memory (LSTM) network was invented and targeted to solve the gradient disappearance problem, the LSTM actually replaces a neuron in the Hidden Layer of the RNN with a more complex structure called memory block (Sherstinsky, 2020). In related work, RNN-based LSTM was taken into account for sentence generation and sentiment analysis. A generative adversarial network was applied to generate tagged sentences for data augmentation and an RNN to extract features from time series data. The findings of the study show the effectiveness of the category sentence generative adversarial network (CS-GAN) (Li et al., 2018).

RNN has also been applied to generate social media text. The exploration of NLG was based on the maturity of NLP techniques and could be used to generate relatively random social media texts. The main RNNs have also been used to generate social media text. The exploration of NLG was based on the maturity of NLP techniques and could be employed to generate relatively random social media texts. The main contributions were the RNNs for linguistic feature extraction and the use of LSTM to solve the learning

dependency problem, as well as an attention mechanism, which proved that the model scores were higher than the HMM model (Cao & Wang, 2018).

2.5 Hidden Markov Model

Hidden Markov Model (HMM) (Baum, 1968) is a simple dynamic probability model, which is mainly used for data modeling of time series. At the same time, there are many practices in natural language processing and speech recognition. Hidden Markov Models are probabilistic models about time series, and the HMM deals with problems that have data with time series information. There are two terms that are key in the HMM. One is the state sequence $\{y_1, y_2 \dots y_n\}, y_i \in Y$ and one is the observation sequence $\{x_1, x_2, x_3 \dots, x_n\}, x_i \in X$. The state sequence randomly generated by a hidden Markov chain is called the state sequence; each state generates an observation, and the resulting random sequence of observations is called the observation sequence. The task of extracting important fields from the titles of computer science research papers using HMMs was 92.9% accurate (Seymore et al., 1999).

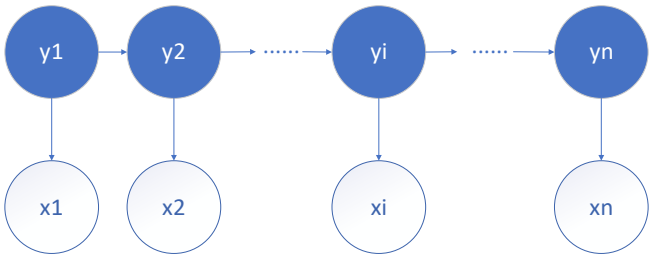

Fig 2.2: Hidden Markov Model

The HMM can also be used on the data collected from user feedback sequences to automatically learn the potential context for each user to predict personalized recommendations. The user preferences that changed over time would affect the accuracy of the recommendation system. Therefore, changes in time, physical, social and emotional aspects could be obtained from click preferences with time stamps collected, thus making the model more accurate (Aghdam, 2019).

2.6 Image caption

The next step is to examine and summaries traditional image captions. This is because there are similarities between storytelling and image captioning in that both require the analysis of images and the generation of textual descriptions based on the content of the images. Image or visual captioning is also a method of describing images based on natural language descriptions using algorithms that use images as input. It often requires the integration of computer vision and natural language processing and is an interdisciplinary challenge. It usually requires the use of vector-to-sequence learning which includes convolutional neural networks, recurrent neural networks, and attention mechanisms.

Generally speaking, existing image subtitle algorithms can be divided into three categories according to the way of sentence generation: template-based method, transit-based method and neural network-based method (Jia et al, 2015). And common evaluation metrics include BLEU, METEOR, CIDEr and SPICE (He & Deng, 2017).

CNNs based on Neural Image Caption (NIC) model are evaluated, the second is based on soft-attention framework. The large dataset takes use of the information extracted from the CNN and then applies frameworks such as decoder encoders to generate sentences in captioning. The image captioning is also employed to provide assistance to people with visual impairments. The study suggests that ResNet and DenseNet are very suitable for generating image captions with low model complexity. In order to generate captions with diversity, model assembly can be used, integrating several different algorithms in one task, for example, using a CNN + LSTM + attention model. The CNN is applied to convert images to one-dimensional vectors and extract visual features, the image features are mapped to the vector space of hidden state of the LSTM， and image feature vectors at each time-step are employed to calculate attention (Katiyar & Borgohain, 2021).

Two shortcomings of the existing CNN-RNN framework method are identified, specifically, the importance of words is not reflected, and semantic objects or scenes are not properly and reasonably combined. Therefore, a new method based on LSTM is proposed to generate a more relevant and accurate description of the image (Ding et al.,

2019). In eq. (2.2), θ^* is the parameter of our model and $\mathcal{L}(\bullet)$ is a pre-defined likelihood function, N is number of training images and M is number of training description sentences.

$$\theta^* = arg\, \max_{\theta} \sum_{m=1}^{M_n} \mathcal{L}\,(S_{nm}|I_n;\theta) \tag{2.2}$$

In eq.(2.3), $\mathcal{O}(\bullet)$ is a pre-defined objective function, this function is use to generate sentence for image J.

$$R = \arg \max_{R'} \mathcal{O}(R'|J;\theta^*) \tag{2.3}$$

The experimental results show that the new method can effectively reduce the error rate of sentences. A global-local attention (GLA) method was proposed, the attention mechanism was proposed to integrate local representation at object-level with global representation at image-level (Li et al., 2017). CNNs are applied to extract global feature and local features of an input image, attention mechanism was employed to integrate local features with global feature, and finally LSTM is utilized to generate the sentence to describe the content of the input image. This method provides a more relevant and coherent natural language sentences result.

Most of the methods adopt the combination of CNN and/or RNN and LSTM to achieve image captioning. But there is another method that uses CNN, LSTM to combination with 11 attention methods to build the model. A midge system-based approach combines probability distributions between words to train sentence structure from caption, using keywords detected from images, lowering prior assumptions about sentence structure (Wang et al., 2020). Describing a picture in sentences has more challenging than words, because it is often difficult to predict the relationships between different objects (Gupta & Davis, 2008).

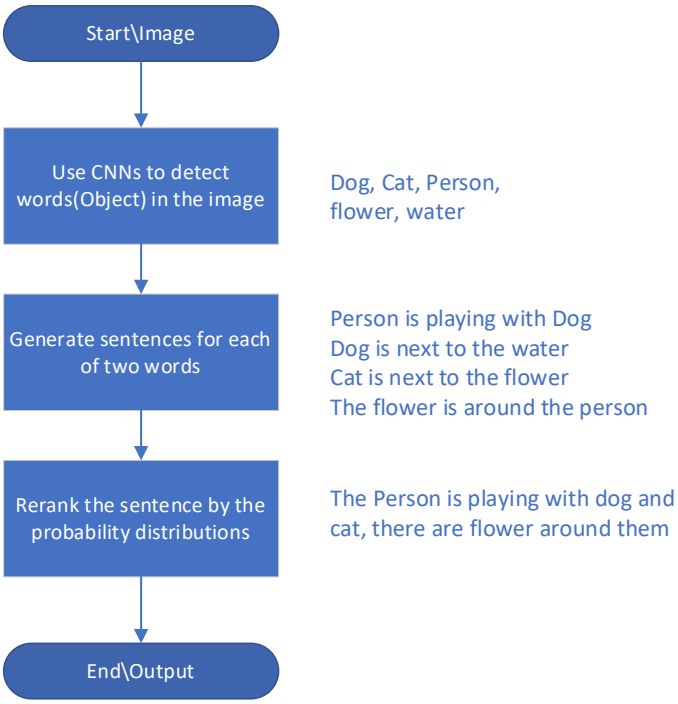

Fig 2.3: The process of using Midge

The algorithmic nature of attention mechanisms is to summarize how attention mechanisms are applied. It is also proposed that for validation methods, it is still a challenge to scientifically evaluate the quality of automatically generated texts because the understanding of natural language is a subjective evaluation for humans (Wang et al., 2020).

Another way is to extract a named entity from an image and fill it with a template, which extracted named entities that appeared in the sport news, generated template captions, and marked placeholders to indicate the need to fill named entities. It then selects the correct named entity with the help of sentence attention (Biten et al., 2019), where I is the input image, I_t the attended image features and I_f are features of the input image extracted from CNN which is ResNet, where h_{t-1} is the hidden state at time $t-1$.

$$I_f = CNN(I)$$
$$I_t = Att(h_{t-1}, I_f) \quad (2.4)$$

The experimental results show that it is effective to use the attentional mechanism to select context based on named entity. Similar practice of using named entities also appeared in Jing's research. Compared with this, they used the global

attention mechanism to better relate the context information of pictures (Jing et al., 2020).

2.7 Object Detection

Target detection differs in some other ways from the usual image classification. Image classification takes an image as input and outputs the corresponding label for the image after analysis. For target detection, however, the task results in the need to find and label a specific object within the image, this often involves localization and detection. This process is also known as extracting feature regions, and common methods include exhaustive search, segmentation, probabilistic models, deep learning (Ding et al., 2017).

By summarizing the research in recent years, we can find that there are two kinds of object detection methods at present, they are One-stage and Two-stage object detection. As milestones of object detection, the representative ones include Region with CNN feature and YOLO (Zou et al., 2019).

The RCNN for object detection was firstly proposed in 2014 (Girshick et al., 2014; Girshick et al., 2015). The Region Proposal network is proposed to implement the object detection problem. The algorithm was divided into three steps: (1) Candidate region selection, (2) CNN feature extraction, (3) Classification and boundary regression. Region Proposal is a traditional region extraction method that generates approximately 2,000 candidate regions based on the selective search method to obtain different image blocks on the current input image that may contain the target, and then crops the image blocks to a fixed size and inputs them into the CNN network to make a determination of the current image block category. In order to obtain more accurate positioning, Regression target T of a bounding box regression model was trained with a training sample of (P, G), where $P = (P_x, P_y, P_w, P_h)$ is the candidate region, $G = (G_x, G_y, G_w, G_h)$ is the real box, and G is the real box with the largest IoU of P (IoU>=0.6).

$$t_x = \frac{G_x - P_x}{P_w}$$

$$t_y = \frac{G_y - P_y}{P_h}$$

$$t_w = \log\left(\frac{\mathrm{G}_w}{\mathrm{P}_w}\right)$$

$$t_h = \log\left(\frac{\mathrm{G}_h}{\mathrm{P}_h}\right) \quad (2.5)$$

As new methods, Faster R-CNN was followed. In 2015, a faster detector Fast R-CNN (Girshick, 2015), which is a method integrating with R-CNN and SPPNet (He et al., 2015) together. Compared with RCNN, the accuracy is improved and the speed is increased by 200 times, which was followed by Faster R-CNN (Ren et al., 2015), a regional proposal network (RPN) was compared to Fast R-CNN.

Compared with the two-stage detection, a completely new one-stage detector is proposed. YOLO was first proposed by Redmon et al. in 2015. Unlike previous proposal detection + validation methods, YOLO take use of a single neural network to act on the entire image (Redmon et al., 2015). YOLO divides the image into multiple regions and simultaneously predicts the probability that the bounding boxes of each region have been classified. YOLO segments an image into S×S grid cells. If the center of an object is in this grid, the grid is responsible for predicting the object. Each bounding box needs to predict B bounding boxes, and each bounding box needs to regression its own position and also need to predict a confidence score.

$$Confidence\ score = \Pr(object) \times IoU_{pred}^{truth} \quad (2.6)$$

If any object falls in a grid cell, the first item that the probability of the bounding box contains the object $\Pr(object)$ is set to 1.00; otherwise, the first item is set to 0. The second item is the IoU value, the accuracy of the bounding box IoU_{pred}^{truth}, which is the predicted bounding box between the actual ground truth. There are five values (*x, y*, *w*, *h*) and confidence are predicted for each bounding box. Then, there are S×S grids, and each grid needs to predict B bounding boxes and C categories. The output is going to be a tensor of $S \times S \times (5 \times B + C)$. Therefore, since the output layer has been a fully connected layer, the YOLO model only supports the same input resolution as the training image during detection. However, in subsequent updates, both YOLOv3 (Jocher et al., 2021) and the latest YOLOv5 (Jocher et al., 2021) have created new milestones that have been greatly improved in speed and accuracy, and the accuracy of identifying small

objects has been improved.

Chapter 3
Methodology

The main content of this chapter is to clearly explain the research methods, which satisfy the objectives of this report. The chapter mainly covers the details of research methodology for storytelling on tourist photo based on object detection using deep learning which will be clearly introduced with the confident and imaginative use of the feature description methods.

3.1 Data Collection and Experimental Environment

Images containing visual objects are employed to train specific deep learning models, and part-of-speech Tagging is supplied to analyze the emotional information contained in a description of a target. By comparing the models generated by CNN and Transformer after learning the same samples, the issues in this scenario are discussed.

For data collection, the famous tourist sites in Nanjing China, such as Nanjing Confucius Temple, ancient Qinhuai, and Qinhuai River, were selected as the main identification objects, which collected the transportation that people usually take to these scenic spots, such as subway. The goal is to enrich the variety of stories that are generated by connecting multiple pieces of information together. The more information provided, the richer the content will be, it looks much like human writing documents.

As a tourist, a lot of photos or short videos were taken during the journey of a day. When making data annotation, the videos were parsed into pictures by extracting frames. The same target from different angles was taken for the purpose of simulating the habits of different tourists. In this way, samples were acquired quickly and the blurred samples were included, which enhance the robustness of the model. For each picture, manual annotation is carried out to make the sample as accurate as possible.

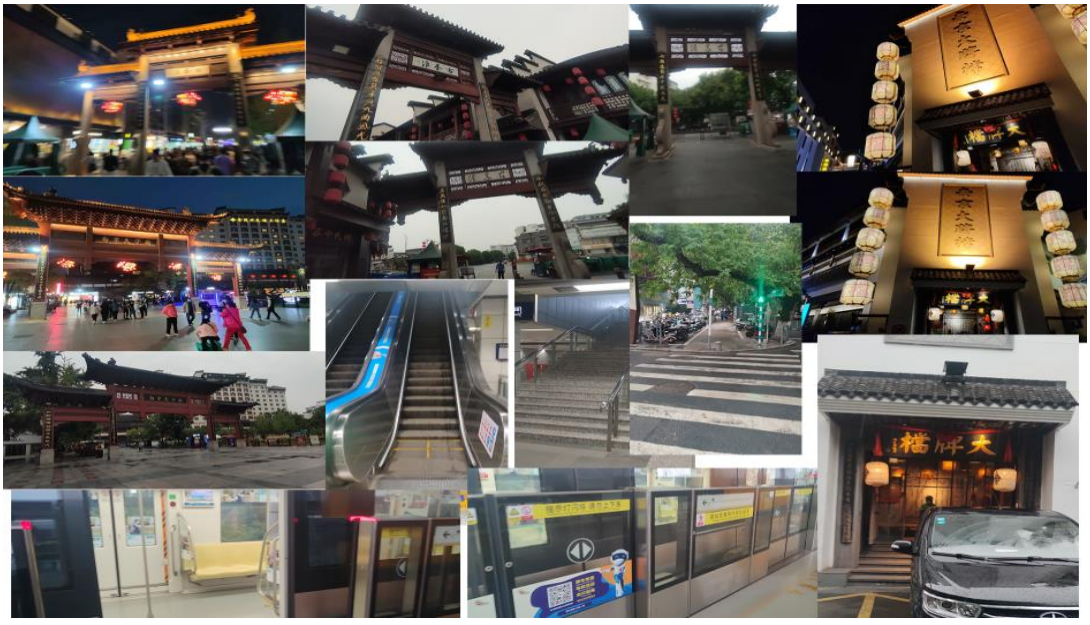

Figure 3.1: The samples of scenic spots

In this project, we use a GeForce 1060 graphics card, a CUDA GPU was ultilized

to speed up the training and test of computations. Ubuntu 20.04 LTS open-source operating system was taken to create a framework based on Anaconda 3 by using Python 3.8.

3.2 Convolutional Neural Networks

CNN is a type of feedforward neural networks. Different from ordinary feedforward fully connected networks, convolution operation can better extract the regional features of the matrix which has better applications in image processing. If the image is input as a fully connected layer, it will be stretched into one-dimensional data and the relationship between surrounding pixels will be lost. Moreover, too many parameters will also lead to the problem of slow calculation. CNNs are similar to deep neural networks that are composed of neurons with learnable weights and bias constants. Each neuron has its input, the output is the confidence of each classification. The convolution function is written as Fig.3.2, given a convolution matrix *K* for the input image.

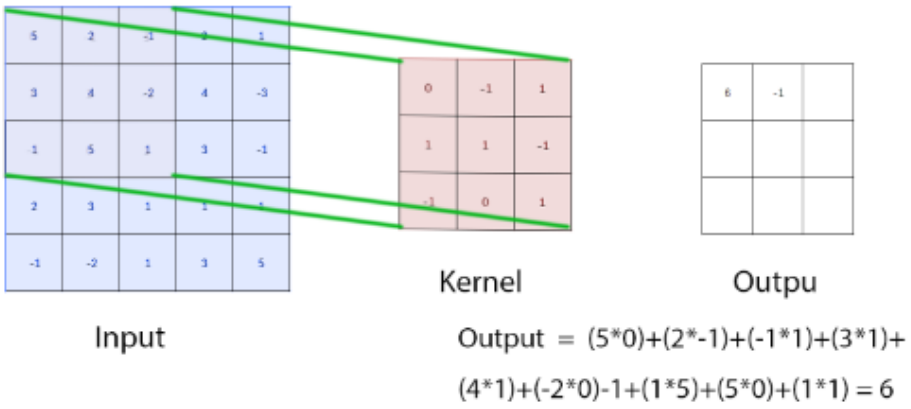

Fig.3.2: CNN process

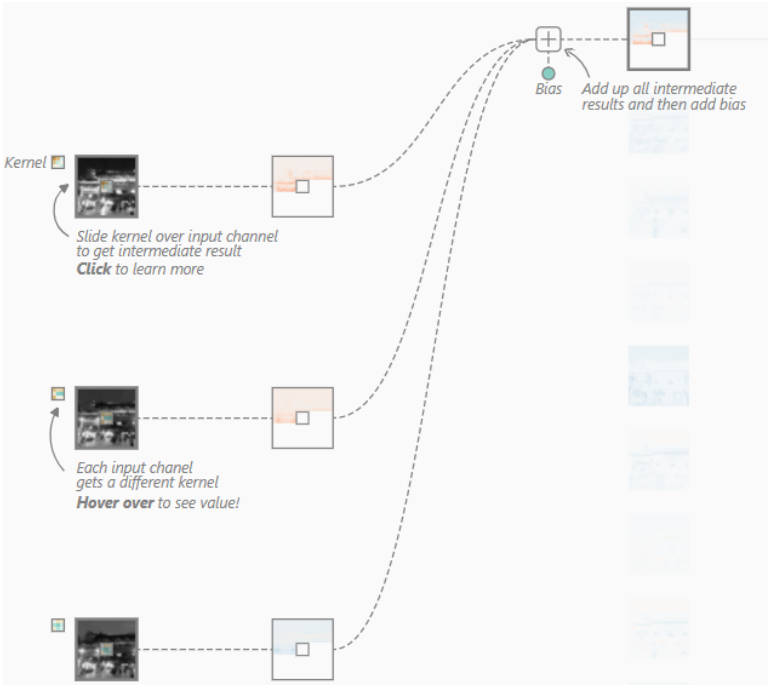

Fig.3.3: The network structure

The calculation is usually conducted by using python library torch.nn, and using nn.Conv2d. In this experiment, pooling layers are applied after multiple convolution. Pooling is also a common operation in the convolution process. It is a subsampled technique with the purpose of reducing feature dimensions. Visual features in a region can be sampled quickly through pooling layer, and the robustness of the model against translation or rotation can be enhanced at the same time, because the output value is calculated by values within the range. Max pooling was chosen for this project, with the goal of retaining more texture information in the model.

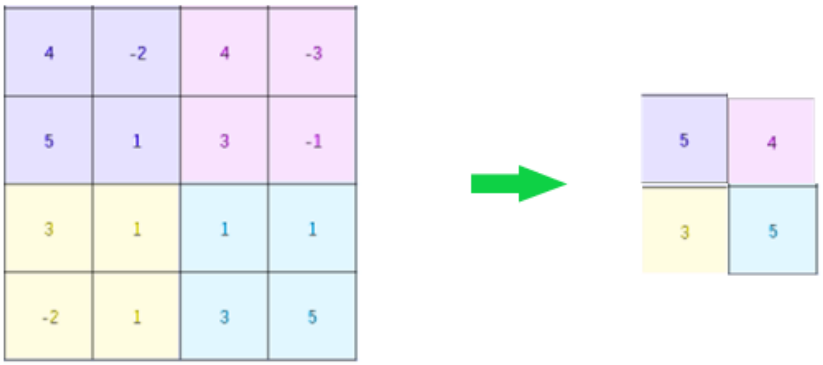

Fig. 3.4: Convolution operation

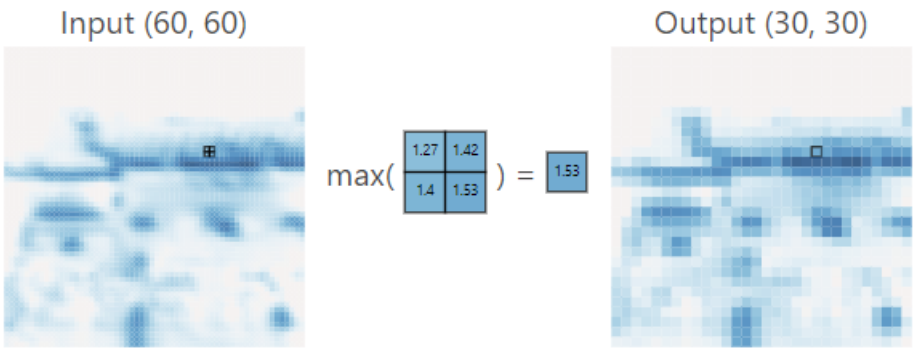

Fig.3.5: Sampling operation

Usually, after the convolution operation and before the full connection layer, we will have the activation function. Throughout continuous learning, we solve the problems that the linear model cannot solve. The SiLU function has been taken into consideration in this experiment, SiLU outperformed than ReLU (Elfwing et al., 2018) as shown in eq.(3.1).

$$a_k(\mathrm{s}) = \mathrm{z}_k \alpha\,(\mathrm{z}_k) \tag{3.1}$$

where in AK SiLU, *s* is input vector, z_k is input to hidden unit *k*. And the input to the hidden layers Z_k is given by

$$z_k(\mathrm{s}) = \sum w_{ik} s_i + b_k \tag{3.2}$$

where $b_k$is the bias and $w_{ik}$is the weight connecting to the hidden units *k* respectively.

After iterative times of convolutions and fusions, the feature map of the image will be expanded into a vector, the feature map vector will be employed as the output, so as to obtain different CNN classification features.

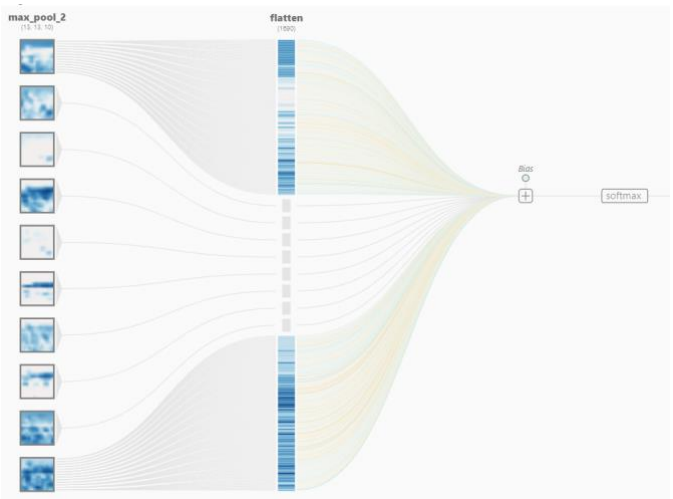

Fig. 3.6: Deep learning-based classification

3.3 Transformer & Attention

In the process of literature review, Transformers have been explained for multiple times as an innovative way of target detection in image captioning (Ding et al., 2019; Li et al., 2017). Transformer is mainly applied to the field of NLP in the early stage and achieved great success. The reason is that Transformer is able to transform words into vectors and rely on the relationship between words in the processing, weakening the influence of the position of words in the text sequence. The Transformers rely on the attention model, studying the relationships between vectors to determine which ones should be given more attention. In the process of image processing, if the image is segmented into multiple vectors, in order to retain the location space information of the patch in the image, it should also be noted that the location information needs to be encoded.

For the attention mode, it is essentially to convert the input image into a vector, and then translate it into a data structure <*Key*, *Value*>. By calculating the weight coefficient between *Key* and *Value*, *Value* is weighted and summed to get the final attention. That is, Query and Key are applied to calculate the weight coefficient of the corresponding which take use of the function as shown in eq.(3.3), t is time or sequence and d_k is hyper-parameter.

$$Attention\ (Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.3}$$

Multihead attention is like the number of X self-attention ensemble. Similar to convolution, convolution uses convolution kernel, but Multihead Attention takes adavantage of multiple self-attention. Finally, multiple features are splicing together by using full connection layer. Transformer replaces RNN with attention. If RNN is trained, the calculation of the current step depends on the hidden state of the previous step, which means it is a sequential procedure, if performing each calculation needs to wait after the previous calculation is completed. As an innovation, Transformer does not use RNN, each calculation will be carried out in parallel, which could significantly improve the speed of training.

The hybrid structure enables CNN to help Vision Transformer (ViT) (Dosovitskiy et al., 2020) get better bias capability, the input sequence is obtained by simply flattening

the spatial dimensions of the feature map and projecting to the Transformer dimension. Therefor in Python, the transformer module is used to replace the C3 layer in the last layer of the backbone.

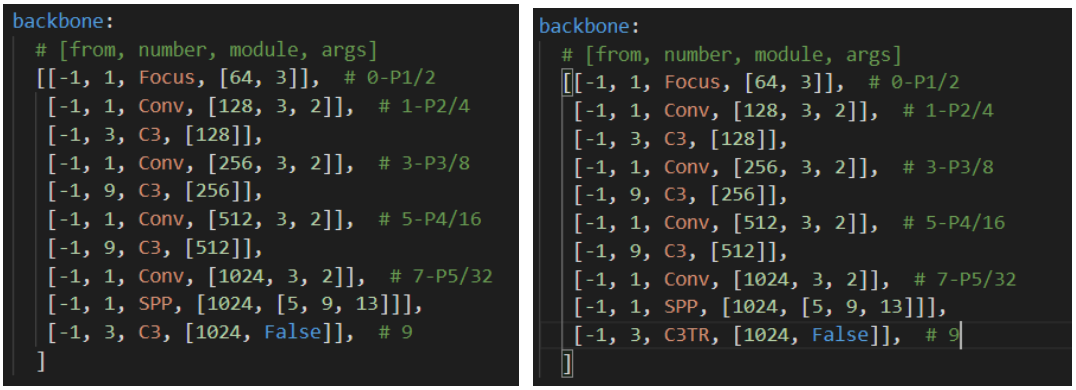

Fig. 3.7: The YOLOv5 & YOLOv5-Transformer

3.4 Object Detection

In traditional object detection system, Deformable Parts Models (DPM) method is adopted to propose target region by sliding frame method, and then classifier is adopted to realize recognition. In this experiment ,we take advantage of the one-stage detection method and directly output the scores and regression for anchors. The advantage of this is that it saves a lot of time and a lot of unnecessary calculations.

Cross Stage Partial DenseNet (CSPNet) is actually based on the idea of DensNet to isolate the feature maps of the base layer by copying the feature maps of the base layer and sending the copy to the next stage via the Dense Block. CSPNet solves the problem of repeated gradient information in network optimization in Backbone of other large convolutional neural network framework, reducing the number of model parameters and FLOPS, which not only ensures the inference speed and accuracy, but also reduces the model size (Wang et al., 2020).

The image is divided into S×S grid cells. If the center of a target falls into a cell, the cell is responsible for detecting the object. Bounding boxes (BBox) and confidence score of each grid cell are predicted. The confidence value represents the confidence that box contains a target. Then, we define the confidence value as zero. If there is no target, the confidence value is zero. In addition, we hope that the predicted confidence value is the

same as the ground truth intersection over union (IOU). Each grid cell predicts the conditional probability value. The probability value represents the probability that the grid contains a target, and each grid predicts only one type of probability. During the test, each box is multiplied by the category probability and box confidence to get a specific category confidence score. This score represents the probability of the category appearing in the box and the compatibility between the box and the target.

3.5 Story Templates

In order to present a more realistic travel diary, we use a template to generate the story, which has the advantage of avoiding stiff sentences. This project will take use of two different templates. One is to take over complete travel routes that conform to the majority of people in a day or within a period of time. The collection of big data is collected and sorted out to fully describe the whole journey without being bored, it is convenient for later modification. There are also narrative stitching templates such as link each photo together according to the timeline and events that took place. The contextual connections between sentences make them hold short-term memories and make the content more realistic. Splicing stories require attention to the handling of turning points, such as taking transport to a place, not to travel on transport. We extend the hidden meaning according to the attributes of the object, for example, the subway is convenient and fast, and the bus is leisurely to enjoy the city scenery. In addition, it will generate sentences based on position relation according to the position relation of the target object by analyzing the picture. We hope to provide safer travel services for the blind through this experiment.

3.6 Training Data Preparation

In order to prevent overfitting of the model due to a small number of samples, the dropout method will be used to randomly abandon some neural units in this study, because dropout makes certain two neurons not always appear in the same subnetwork structure. Based on the weight of the update is not dependent on a fixed relationship between the implicit node work together, and to let the model learn how to extract certain information under

the condition of random neural unit missing, reduce the parameters to prevent model rely on training data too much, and increase the parameters on the generalization ability of a data set, which can improve robustness of the model.

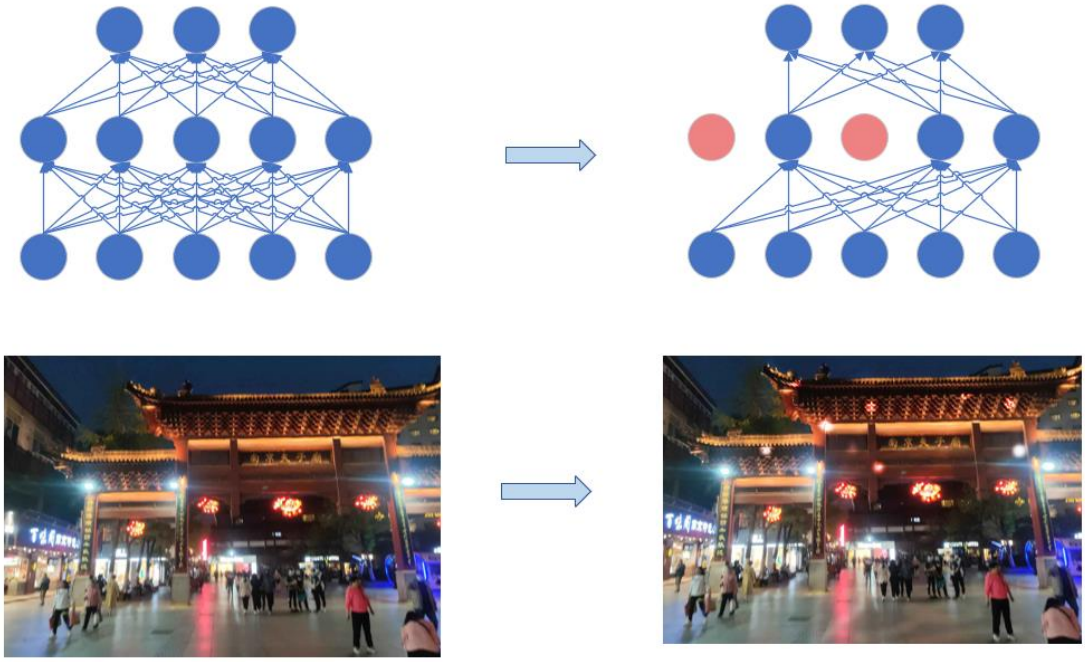

Fig. 3.8: From deep nets to applications

For our model, all samples will be scaled to a uniform size and then input into the neural network. In order to expand the diversity of samples, the samples will be randomly clipped and scaled. Namely the image under the condition of the same size, some samples are fill the canvas or some of narrow to center, and can rotate to within a certain range with random samples, the purpose is to adapt to different visitors of habit, although the majority of cases can be clearly won't appear upside down, but for the photos of individuation, or may appear tilt goal, doing this so can also improve the generalization ability of the model.

Fig. 3.9: Images with rotating, cropping, and padding operations

HSV(i.e., Hue, Saturation, Value) is a color space created by A. R. Smith in 1978 based on the intuitive characteristics of colors. By randomly adjusting hue, saturation and value, the data is expanded to simulate the color difference caused by cameras of different brands or mobile phone cameras shooting the same target. It can also be utlized to simulate different white balance effects in different weather scenes or to simulate additional filters added by users. The effect of this is to minimize the generalization ability caused by the lack of a large number of samples and simulate as many colors and textures as possible, which we call it color space adjustment.

Fig. 3.10: Images with various color spaces

3.7 Evaluation Methods

In order to evaluate the reliability of the model, we firstly need to clarify the understanding of classification accuracy.TP, TN, FP, and FN are mainly used to describe the four types of problems in binary classification, as shown in Fig. 3.11, based on these concepts, precision and recall will be explained.

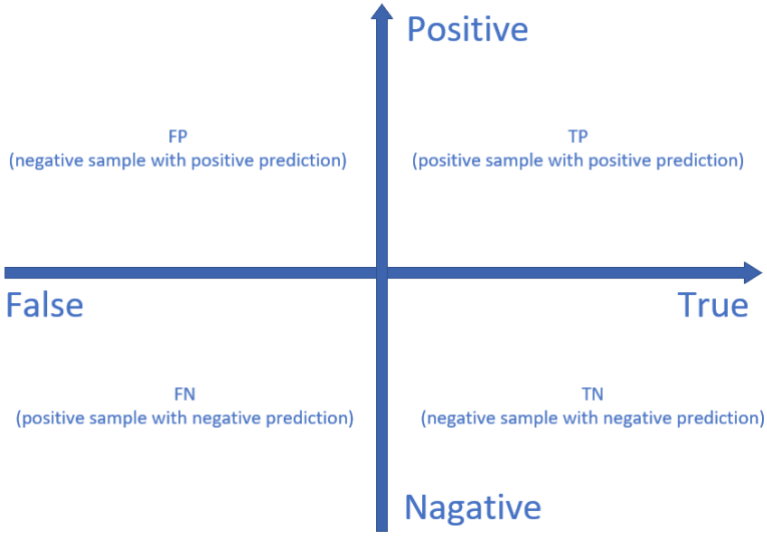

Fig. 3.11: Classification diagram

Precision measures the probability of positive samples that a classifier classified is indeed positive.

$$Precision = \frac{TP}{TP+FP} \tag{3.4}$$

But it needs to consider that, if the accuracy is 100.00%, all positive classes from the classifier are indeed positive classes. If the accuracy is 0.00%, it means that none of the positive classes in the classifier are positive. Precision alone doesn't measure how good a classifier is, given a case that with 50 positive samples and 50 negative samples. The classifier classifies 49 positive samples and 50 negative samples into negative samples, the remaining positive samples into positive samples.

Recall means measures the ability of a classification to find all positive classes.

$$Recall = \frac{TP}{TP+FN} \tag{3.5}$$

The IOU is also used to determine the TP of the samples. IOU calculates the ratio of Intersection Over Union (IOU) of the predicted bounding box and the real bounding box.

IOU is a very important function in performance *mAP* calculation in object detection algorithm. In order to calculate the IOU, it needs to calculate the intersection which is the area of overlap, and the sum of the two areas that only contain one intersection area as showin in eq. (3.6)

$$IOU = \frac{Ground\ truth \cap Predict}{Ground\ truth \cup Predict} \tag{3.6}$$

The measurement of recognition accuracy in object detection is called mean average precision (*mAP*). In multiple classes of visual object detection, each class can draw a curve based on recall and precision. *AP* is the area under the curve, *mAP* means averaging the AP of each category.

$$AP = \int_0^1 p(r)dr \tag{3.7}$$

For the verification of the generated story, it mainly depends on the accuracy of target detection. Because we chose template-based sentence generation in this experiment, it can be inferred that of the target detection accuracy is improved, the generated sentence will be more accurate.

3.8 Training Program Implementation

The purpose of this research is to develop a mobile phone Apps so that tourists can quickly analyze the photos taken by mobile phones and obtain the stories generated according to the photos. Uploading photos on your phone not only allows you to use image resources, but also the EXIF information that is unique in photo files, and help to get more information and generate stories.

This project is mainly divided into back-end and front-end. The back-end is mainly used for material collection, model training, detection and return detection results after receiving pictures. Because of Python's easy to get started simplicity and the availability of plentiful, complete third-party open-source materials, this study uses Python to achieve back-end development goals.

Thanks to the generous open-source community environment, this study will refer

to YOLOv5 as the basic research framework to discuss the effect of YOLOv5 in the object detection of tourist attractions, and will also introduce Transformer and attention model into CNN to optimize the network. Firstly, YOLOv5s was selected for a clean initial training of the model and this study will introduced Transformer Layer as the last layer in Backbone, and allow it to connect with other features later on. Since this experiment did not contain small targets, there was no need to consider the case of small object. Although in the vast majority of cases, a tourist photo almost should be positive, it does not rule out that tourists may take photos of various strange angles for artistic reasons, so the need for random rotation is also necessary.

Color adjustment was introduved earlier, because the collection only from two mobile phones, one for training, the other is for the validation material, which is likely to avoid use of the same cell phone to provide both training materials, and provide verification material. In the case of a small number of samples, the random adjustment for colors is meaningful, this can enhance the robustness of the model. As an experiment, the study will also examine the effect of using pure black and white photos with the original photos on a same dataset.

As part of this research work, we hope to find unique experimental data. Even though YOLOv5 shows the detected object directly, a part of the code still needs to be added and modified to meet the requirement of the project according to the objectives of the study. We modified detect.py, added a new method that allows external custom parameters to be passed in, and now it is an option to save the detection results locally according to the task scheduling instructions, package the results into JSON and send them back over the network. In the part of network request processing, Python is also used to build the proposed LAN structure, so that the simulator running locally can access the service request interface which created and run by Python.

For the client part, a basic demo will run on any popular Android phone. A simple interface is designed to facilitate users to upload images. After received the images at the server side, relevant results will be returned and a story will be generated automatically to read for users. The story is easily edited manually. At the same time, the mobile terminal can also collect and annotate data, which is convenient to provide more new

samples for the training of the back-end neural network.

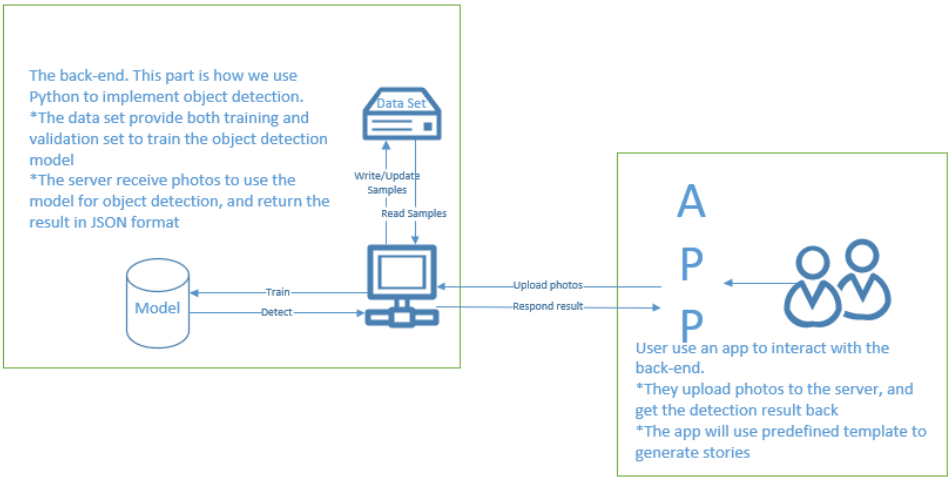

Fig. 3.12: Our proposed model

3.9 NCNN Implementation

For a future scenario that we hope to explore, we hope that through mobile devices, such as mobile phones, or portable raspberry PI, technology can better serve the disabled and help the blind to avoid the risks brought by going out as much as possible. NCNN is a high-performance neural network reasoning framework optimized for mobile platforms. It supports most commonly used CNN networks, including Classical CNN, light-weight CNN and Face Detection, based on NCNN. The researchers were able to easily port deep learning algorithms to mobile devices for efficient execution.

Chapter 4

Results

The main content of this chapter is to show different data, including training set and demonstrate the experimental results.

4.1 Object Detection Result

First of all, in this project we defined the classes of data sets used in training model and evaluation process, the training data and test data were mixed with photo taken with different days and use two different phones.

Table 4.1 Trip data statistics

	Confucius temple	Ancient Qinhuai	Nanjing food stall	Pingjiang bridge	Roast duck	Osmanthus cake	Duck blood soup	Total
Training set	300	280	190	55	80	80	120	1065
Validation set	50	50	40	10	20	20	35	225

Generally speaking, on the training object detection model result, the object detection model trained by tourism photos has high robustness with detection precision up to 0.9923 mAP, which is considered to be able to effectively identify key object from photos. It is very practical for extracting key objectives and generating tourism stories.

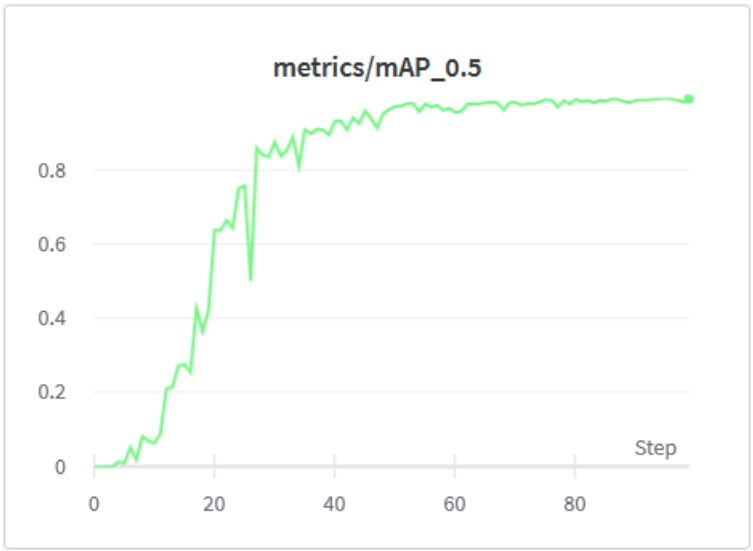

Fig. 4.1: Result of our model mAP@0.5

The sample number was not large enough, so after 80 iterations, the results tend to be stable, and the model might be over-fitting. Although the accuracy and recall rate were close to 1, it might also be because our model did have high robustness. As can be seen from the picture of MAP_0.5:0.95, The accuracy keeps rising, indicating that the boundary boxes predicted by the model still maintain the generalization ability.

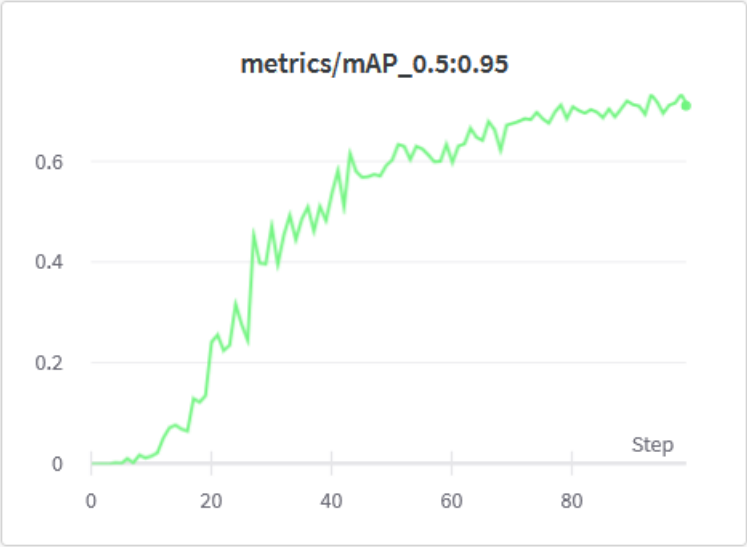

Fig. 4.2: Result of our model mAP@0.5:0.95

YOLO network segment the input image into SxS (S = 19) grid according to the size of feature map (YOLOv5s) (Jocher et al., 2021). If the center of an object falls in a cell, the cell is responsible for detecting the object. The cell will output multiple predicted bounding boxes and the confidence of each predicted bounding boxes. If the bounding boxes with low confidence are discarded, we found that there will be multiple overlapping detection boxes for the same object in the detection process, and non-maximum suppression (NMS) (Neubeck & Van Gool, 2016) will take effects at this time.

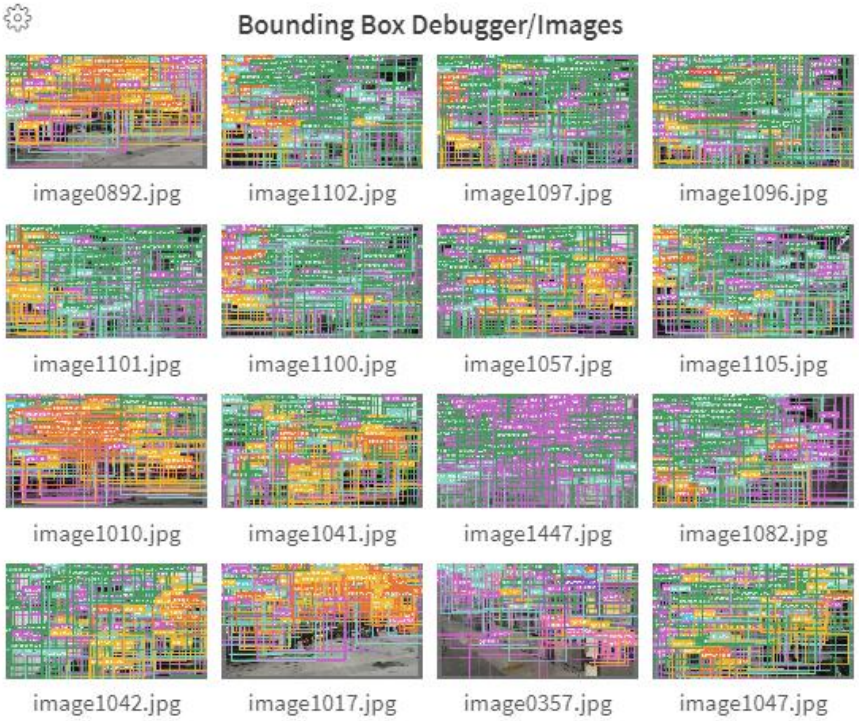

Fig. 4.3: Bounding box debug images @ Initial iteration

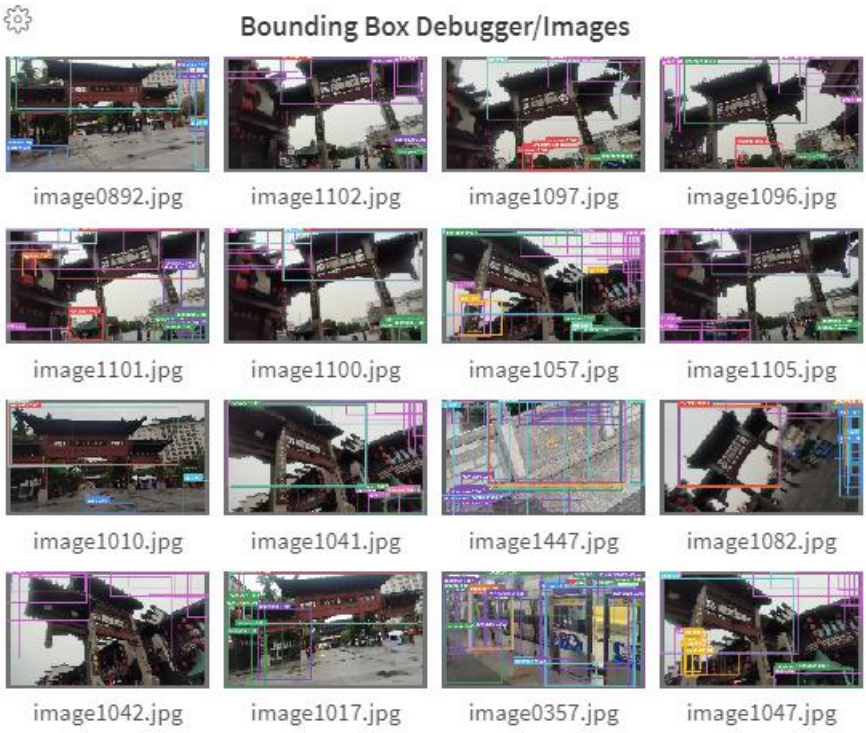

Fig. 4.4: Bounding box debug images @ Final iteration

Our purpose is to eliminate redundant detection bounding boxes in the object detection process according to the preset threshold value. The final position of visual object is located, the output prediction box format is "*xywh*", where (*x*, *y*) represents the center position of the predicted bounding boxes, *w* and *h* indicate the length and width of the predicted bounding boxes. In addition, from the results, the continuous improvement of precision and recall also indicates that the accuracy of positioning gradually tends to be stable with the training process, the NMS reduces the number of repeated candidate boxes with lower confidence.

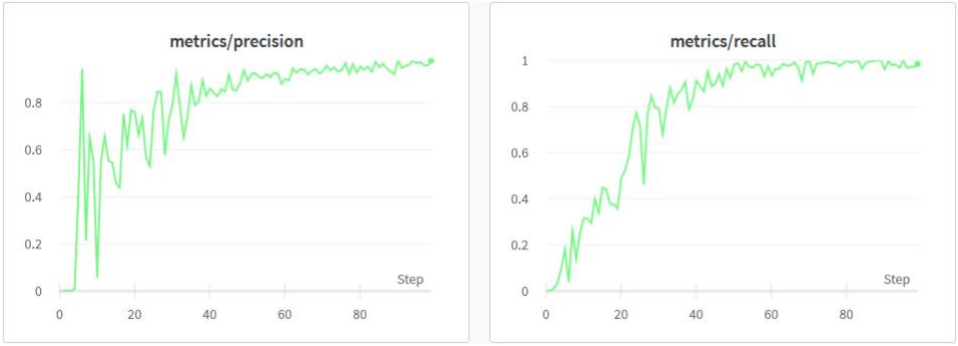

Fig. 4.5: Training result of precision and recall

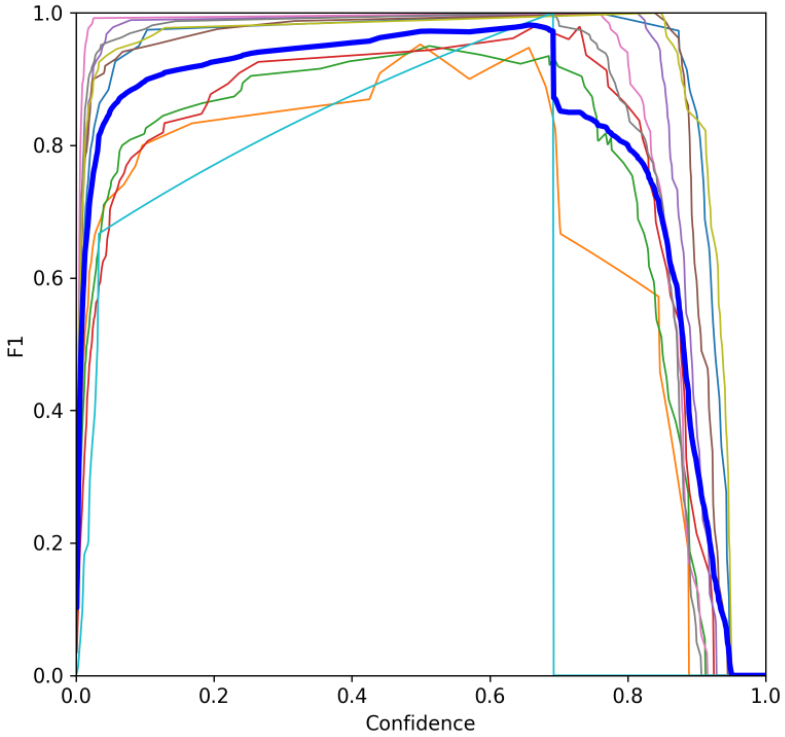

Fig. 4.6: F1 Curve shows the average confidence for every classification

In Fig. 4.6, we see that the accuracy of the model is reliable for each classification, and the graph of F1 Curve can confirm that there is no category with low recognition rate.

4.2 Story Generation Result

In this report, predefined templates are applied to generate stories and fill the templates with identified named entities. Each named entity has multiple templates, if the result of object detection is correct, the generated story will be correct. Moreover, it is difficult for a story to be judged, because there are multiple description methods for the same object, and human language is very subjective (Wang et al., 2020). In the picture, the story has two templates, one is based on the summary of the full journey template, the other is also a narrative travel template based on time lines. We also pay attention to specific elements of the scene, such as whether it's a holiday, whether there are too many tourists, etc.

For example, the three photos on the picture are Confucius Temple, Ancient Qinhuai and Nanjing Food stalls. Because we visited the whole scenic area, the story generated according to the complete module is:

"This trip came to Nanjing, in the evening came to Confucius Temple Qinhuai River

scenic zone, here is one of the most prosperous places in Nanjing, 5A free scenic area. Here you can not only see the historic buildings of Nanjing, the ancient capital, but also eat the most authentic Qinhuai snacks. Go deep into the streets or go boating along the Qinhuai River to experience the local customs and customs of the river from different perspectives. There are many shops on both sides, in which there are many Jinling specialties, such as Yuhua stone, tea, salted duck and cakes, but the price is higher than other places." This method collects a lot of templates through big data, and then generates them directly when the conditions are met. Such as the tourist visited all the related place.

It's a different style to use another way of connecting based on the timeline.

"At about noon, we arrived at The Confucius Temple scenic spot in Nanjing, which was a place of learning palaces, academies and imperial examinations for more than one thousand years in ancient times. Then continue to visit the Qinhuai River, where the road is lined with specialty stores. Finally, I went to Nanjing food stall. There are many special snacks in Nanjing, such as Nanjing roast duck and Osmanthus cake, which are highly recommended."

This method takes use of templates to fill in the blanks, such as when and where the photo was taken, whether the property of the object is a scenic spot or a restaurant, and then uses appropriate connectives.

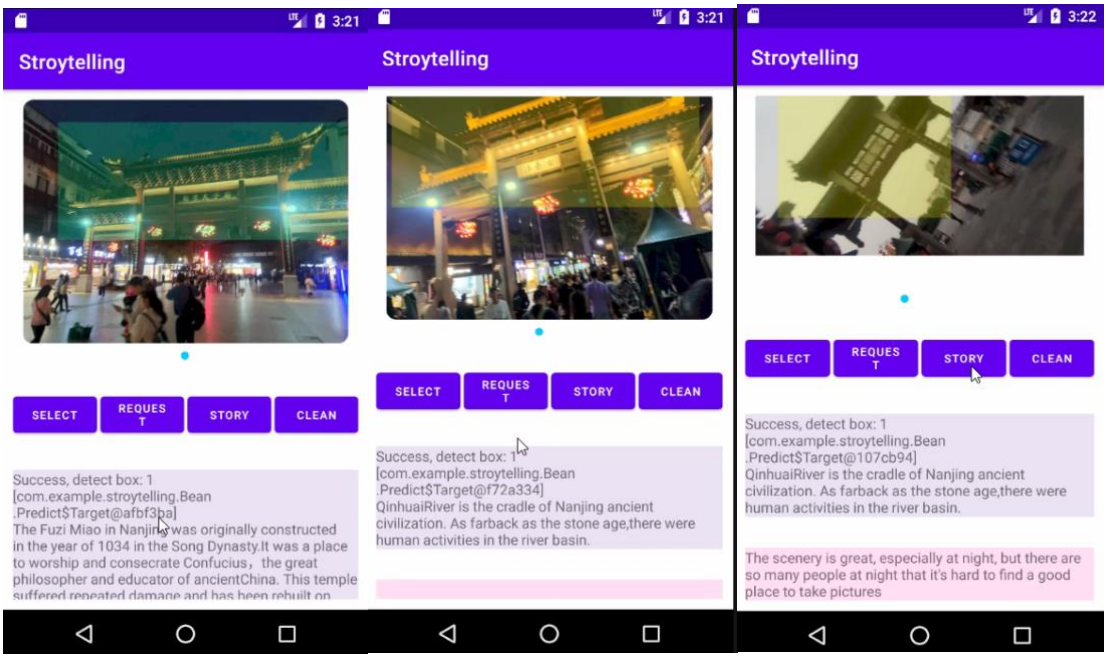

Fig. 4.7: The screenshot of developed application

Through the above experiments, we see that the method of object detection firstly identifies the places tourists have visited, and then fill in the blanks according to the template to generate a travel diary.

In addition, in this paper, we use object detection method to detect the signal lights on the sidewalk, and give corresponding text hints according to the color of the signal lights, such as the red light is not allowed to pass.

Chapter 5
Analysis and Discussions

In this chapter, the research results are summarized and discussed in this chapter. The research results under various conditions will also analysis the result and compare between different methods. The research questions will be answered as well in this chapter. The limitations of this study will also be explained

5.1 Analysis

For object detection, both algorithms and data samples may have an impact on the results and accuracy. Therefore, in this project, we set up a comparative experiment, explore the impact of algorithm selection and samples on the detection of tourist photos. In order to make more comparisons, we selected two new scenarios.

Transformers have new research results in the field of computer vision, but may not produce better results in the case of a small number of samples, so similarly, while training data and parameters remain unchanged, Transformer and attention were added to the last layer of the backbone network. As an attempt to experiment on what impact transform will it have on the traditional convolutional neural network.

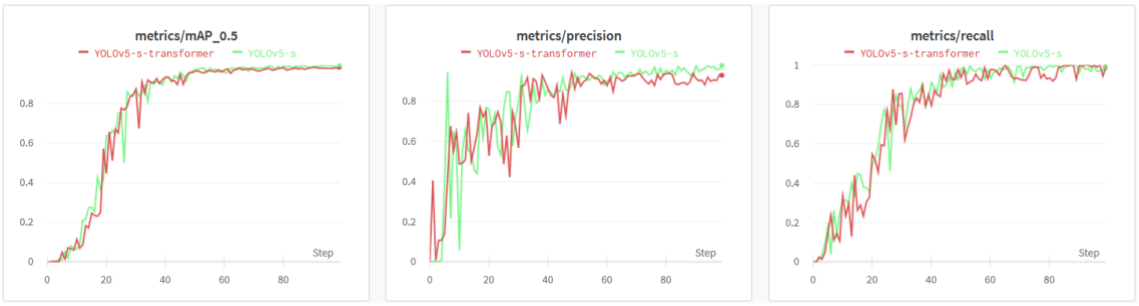

Fig. 5.1: The comparison between YOLOv5-s and YOLOv5-s added transform

Transformers have little influence on the results, the precision rate is lower than that of using CNN completely, which indicates that more negative samples are predicted to be positive samples. The possible reason is the lack of inductive bias, because if this problem is to be solved, more training data samples are needed to alleviate the problem of inductive bias.

As for the recognition of tourist attractions, in the previous experiment that the trained model would overfit due to the lack of enough samples. Therefore, as a hypothesis, in order to test the influence of training results based on the different training dataset, a plethora of comparison schemes were designed. Under the condition, the same color space transformation scheme is maintained during the training process, the photos shot in the daytime are used as the training set and the photos taken at night are used as the verification set to observe the generalization of the model.

Table 5.1: Confucius temple data statistics

	FZM	FZM_Day_0%Night	FZM_Day_10%Night	FZM_Day_20%Night
Daytime Training Set	200	200	200	200
Nighttime Training Set	50	0	5	10
Nighttime Validation Set	50	50	50	50

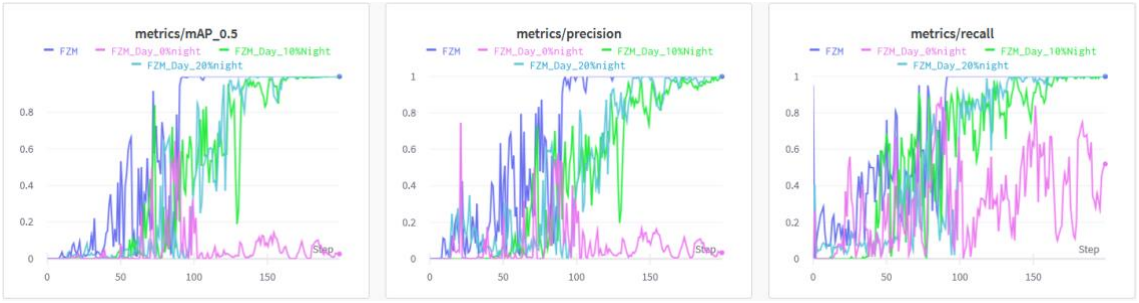

Fig. 5.2: The comparison between adding different numbers of nighttime samples

From the results that the more nighttime samples are added, the better the identification ability of night scenes will be, rather than complete absence of nighttime samples. However, we see that even if a small number of images are added, the generalization ability of the model can be expanded. Therefore, a more varied training set is helpful.

Furthermore, the study compared the colors of the input images. A complete data set was used, but as a control, one set of data was colored and the other was black and white, so that the effect of color on the training model could be explored.

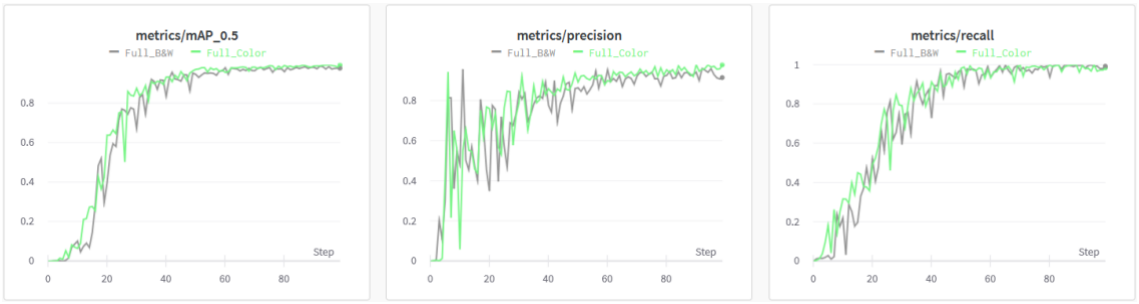

Fig. 5.3: The comparison between colored photo and black and white photo

While switching to black and white didn't have much effect on the results, which were the same in terms of training speed or accuracy. Therefore, it can be found that object boundary and texture are the core of target detection, and color filtering has no impact on

performance and accuracy because RGB conversion to black and white is still different.

For most tourist photos, the object to be detected is usually not too small. Therefore, in previous experiments, the image resolution 640x640 was selected for the input image. In case of accelerated training and recognition, smaller images are also an option. As an assumption, we resize image to 320x320 for training model, so as to explore the correlation on image resolution between detection rate and speed.

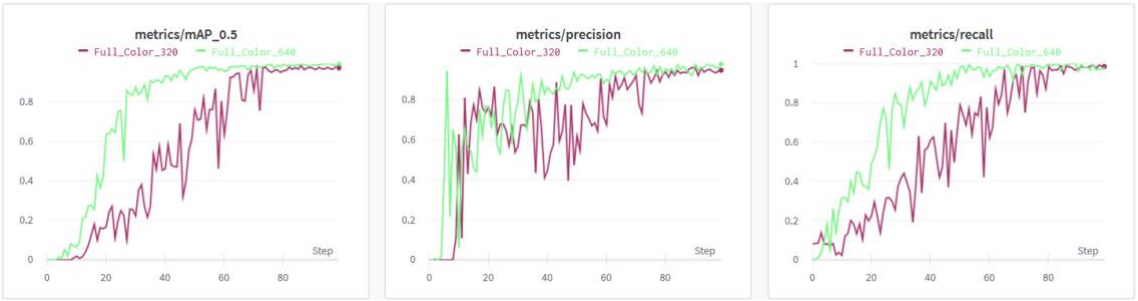

Fig. 5.4: The comparison between different image size

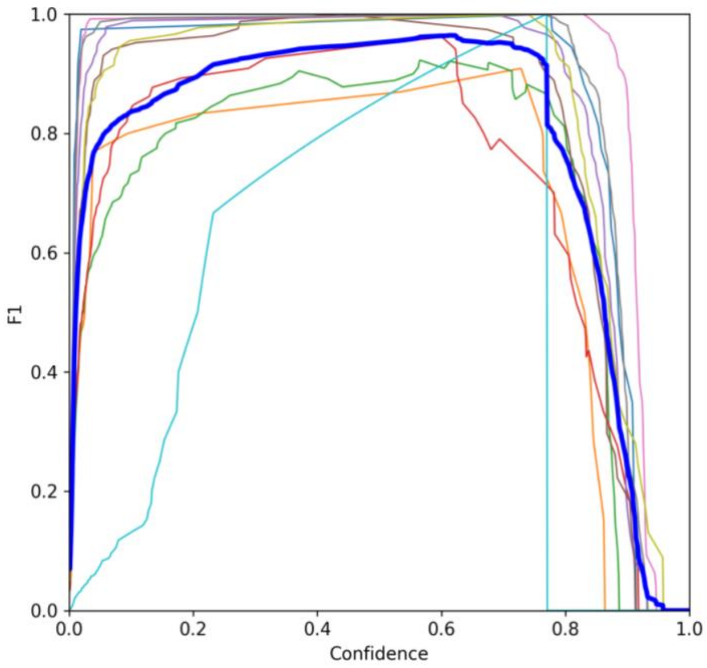

Fig. 5.5: F1 Curve based on 320x320 dataset

From the experimental results, we see that when the image size is reduced from 640×640 to 320×320, Mean Average Precision (mAP) @ 0.5 was achieved as before, but with about 20 epochs later than before, and only half of the time used. However, by comparing F1 curve, we found that there is almost no difference between the accuracy of different classification with 640×640 dataset. Therefore, a smaller model is considered for visual object detection in most scenarios, and then a larger model is employed if there

is no result.

5.2 Discussions

Because the template-based story generation was implemented without image annotation, the discussion of experimental results will be split into two parts: Target detection results and story generation.

In the experiment, visual object detection was applied to identify named entities, traditional CNN and CNN+ Transformers were adopted, and the same dataset was taken to test the performance and accuracy of the two algorithms. Through the analysis of the results, we found that there is no obvious difference between the results of the two algorithms, which may be due to the limitations and the problems of bias and rotation in the case of a small data set for Transformers. It is also found that diversity in the same scene is still important, and only relying on the transformation generated by color space is not enough to improve the generalization of the model. Even if a small number of samples with different colors are added, the generalization ability of the model can be greatly improved. In addition, using a low-resolution model for preliminary detection is a method to improve the training speed and detection speed, which effectively reduced the training and detection cost of the model. Through experiments, it is found that the size of 320x320 can still provide good accuracy.

For generating stories using templates to fill in named entities, this method generates more emotion in the context. It's also useful to add more background to the story, such as weather and introduction of scenic spots. The information about a photo is not just the image itself, but also the Exchangeable Image File Format (EXIF) information that generates the image file. The information can be extracted including GPS positioning, time stamp, focal length and exposure, etc., which is able to improve the accuracy of recognition. For example, GPS positioning information can be used to determine the location of the photo, so as to determine whether the photos are distant between each other. Time stamps determine the order in which a trip was filmed. These can be used to add more detail to the story based on LTSM algorithms in template-based cases. The

addition of this information can make a template-based story much more varied than the description generated by the image caption. The downside is that this approach relies on adding more templates.

5.3 Limitations

In this project, most of the photos were taken up close to the object. There are not much covering relationship, almost all the objects in the photos are complete, so the detection of partially covering samples is not enough. The missing samples are not limited to partial covering, but also include the shooting of samples from a long distance, which may be fuzzy in details, especially after resizing, more details may be lost. More, filters and stickers will also have an impact on the photo, this part of the sample should also be added. Overall, more samples are added to improve the robustness of the model.

For generated stories, there is also a need to consider the generality of templates in cases where they are dependent on them. For example, for different cities or scenic spots with different styles, whether there are more general templates or specific templates.

Chapter 6
Conclusion and Future Work

In this chapter, we summarize the subject and method of this project, and propose new research direction according to the result and insufficiency of the experiment, preparing for the future work.

6.1 Conclusion

The purpose of this report is to propose a travel story generation method based on deep learning, because the research project takes use of visual object detection and story template to generate travel stories. In this report, the methods and principles of common object detection and image captioning are simply explained. In addition, various algorithms and parameters are applied to train the model for the photo samples in the study. The results show that the object detection from digital images based on deep learning is up to 99.23% mAP, the accuracy rate 97.21% was achieved by using smaller models in most scenes. It is proved that object detection can be used to identify scenes and named entities in photos, and fill the named entities is a way of generation travel story.

6.2 Future Work

More ideas are proposed for future exploration of image-based storytelling. Using deep learning to describe images is a valuable research area, which involves not only object detection and natural language processing, but also multi-disciplinary integration to enable computers to understand images. Combining CNN + RNN + Attention is a direction worth exploring, and LSTM can better help neural network to fuse the context other than the image pixel content. What's more, the storytelling combined with text-to-speech can also assist visually impaired patients in their daily life and travel, helping them to explore the world.

References

Aghdam, M. H. (2019). Context-aware recommender systems using hierarchical hidden Markov model. Physica A: Statistical Mechanics and its Applications, 518, 89-98.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In International Conference on Engineering and Technology (ICET) (pp. 1-6). IEEE.

Basha, S. S., Dubey, S. R., Pulabaigari, V., & Mukherjee, S. (2020). Impact of fully connected layers on performance of convolutional neural networks for image classification. Neurocomputing, 378, 112-119.

Baum, L. E., & Sell, G. (1968). Growth transformations for functions on manifolds. Pacific Journal of Mathematics, 27(2), 211-227.

Biten, A. F., Gomez, L., Rusinol, M., & Karatzas, D. (2019). Good news, everyone! context driven entity-aware captioning for news images. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12466-12475).

Cao, J., & Wang, C. (2018). Social media text generation based on neural network model. In International Conference on Computer Science and Artificial Intelligence (pp. 58-61).

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In International Conference on Machine learning (pp. 161-168).

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification? IEEE Transactions on Image Processing, 24(12), 5017-5032.

Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018). Convolutional neural network

(CNN) for image detection and recognition. In International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 278-282). IEEE.

Chollet, F. (2017). Deep Learning with Python. Simon and Schuster.

Copeland, J. (2015). Artificial Intelligence: A Philosophical Introduction. John Wiley & Sons.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems, 2(4), 303-314.

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: A new paradigm to machine learning. Computational Methods in Engineering, 27(4), 1071-1092.

Dhyani, M., & Kumar, R. (2021). An intelligent Chatbot using deep learning with Bidirectional RNN and attention model. Materials Today, 34, 817-824.

Ding, X., Luo, Y., Yu, Q., Li, Q., Cheng, Y., Munnoch, R., Xue, D., & Cai, G. (2017). Indoor object recognition using pre-trained convolutional neural network. In International Conference on Automation and Computing (ICAC), Automation and Computing (ICAC),2017 23rd International Conference On, 1–6.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. arXiv preprint arXiv:1602.07776.

Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks, 107, 3-11.

Girshick, R. (2015). Fast R-CNN. In IEEE International Conference on Computer Vision

(pp. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 580-587).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1), 142-158.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT press.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222-2232.

Guo, T., Dong, J., Li, H., & Gao, Y. (2017). Simple convolutional neural network on image classification. In International Conference on Big Data Analysis (ICBDA) (pp. 721-724). IEEE.

Gupta, A., & Davis, L. S. (2008). Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In European Conference on Computer Vision (pp. 16-29). Springer, Berlin, Heidelberg.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1904-1916.

He, X., & Deng, L. (2017). Deep learning for image-to-text generation: A technical overview. IEEE Signal Processing Magazine, 34(6), 109-116.

Hemdan, E. E. D., Shouman, M. A., & Karar, M. E. (2020). COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images. arXiv preprint arXiv:2003.11055.

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In Neural Networks: Tricks of the Trade (pp. 599-619). Springer, Berlin, Heidelberg.

Hinton, G. E., & Sejnowski, T. J. (1999). Unsupervised Learning: Foundations of Neural Computation. MIT press.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).

Jia, X., Gavves, E., Fernando, B., & Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In IEEE International Conference on Computer Vision (pp. 2407-2415).

Jing, Y., Zhiwei, X., & Guanglai, G. (2020). Context-driven image caption with global semantic relations of the named entities. IEEE Access, 8, 143584-143594.

Jocher, G., Kwon, Y., Guigarfr, perry0418, Veitch-Michaelis, J., Ttayu, … Shead, T. M. (2021). ultralytics/yolov3: v9.5.0 - YOLOv5 v5.0 release compatibility update for YOLOv3 (Version v9.5.0). doi:10.5281/zenodo.4681234

Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, … wanghaoyang0106. (2021). ultralytics/yolov5: v6.0 - YOLOv5n “Nano” models, Roboflow integration, TensorFlow export, OpenCV DNN support (Version v6.0). doi:10.5281/zenodo.5563715

Katiyar, S., & Borgohain, S. K. (2021). Comparative evaluation of CNN architectures for image caption generation. arXiv preprint arXiv:2102.11506.

Ketkar, N., & Santana, E. (2017). Deep Learning with Python. Berkeley, CA: Apress.

Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). A guide to convolutional neural networks for computer vision. Synthesis Lectures on Computer Vision, 8(1), 1-207.

Kim, P. (2017). MATLAB deep learning. Machine Learning, Neural Networks and Artificial Intelligence, 130(21).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Li, L., Tang, S., Deng, L., Zhang, Y., & Tian, Q. (2017). Image caption with global-local attention. In AAAI Conference on Artificial Intelligence.

Mitchell, T. (1997). Machine Learning.

Neapolitan, R. E., & Jiang, X. (2018). Artificial Intelligence: An Introduction to Machine Learning. CRC Press.

Neubeck, A., & Van Gool, L. (2006). Efficient non-maximum suppression. In IEEE International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 850-855)

Oh, Y., Park, S., & Ye, J. C. (2020). Deep learning COVID-19 features on CXR using limited training data sets. IEEE Transactions on Medical Imaging, 39(8), 2688-2700.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. Multim. Tools Appl. 79(27-28): 19925-19944

Pan, C., Liu, J., Yan, W., Cao, F., He, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and Two-Stream hybrid networks. IEEE Trans. Image Process. 30: 4773-4787

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Raschka, S., & Mirjalili, V. (2017). Python machine learning: Machine learning and deep

learning with Python. Scikit-Learn, and TensorFlow. Second edition.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances In Neural Information Processing Systems, 28, 91-99.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach, ebook, global edition.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681.

Seymore, K., McCallum, A., & Rosenfeld, R. (1999). Learning hidden Markov model structure for information extraction. In AAAI-99 Workshop on Machine Learning for Information Extraction (pp. 37-42).

Sharp, M., Ak, R., & Hedberg Jr, T. (2018). A survey of the advancing use and development of machine learning in smart manufacturing. Journal of Manufacturing systems, 48, 170-179.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

Smilevski, M., Lalkovski, I., & Madjarov, G. (2018). Stories for images-in-sequence by using visual and narrative components. In International Conference on Telecommunications (pp. 148-159). Springer.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language

modeling. In Annual Conference of the International Speech Communication Association.

Vedaldi, A., & Lenc, K. (2015). MatConvNet: Convolutional neural networks for MATLAB. In ACM International Conference on Multimedia (pp. 689-692).

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 390-391).

Wang, H., Zhang, Y., & Yu, X. (2020). An overview of image caption generation methods. Computational Intelligence and Neuroscience, 2020.

Yan, W. (2021). Computational Methods for Deep Learning Theoretic, Practice and Applications, Springer.

Yan, W. (2019). Introduction to Intelligent Surveillance – Surveillance Data Capture, Transmission, and Analytics (Third Edition) Springer.

Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. Drug Discovery Today, 22(11), 1680-1685.

Zhou, D. X. (2020). Universality of deep convolutional neural networks. Applied and Computational Harmonic Analysis, 48(2), 787-794.

Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055.