Image-Based Storytelling Using Deep Learning

YULIN ZHU

Auckland University of Technology, Auckland 1010, New Zealand

WEI QI YAN

Auckland University of Technology, Auckland 1010, New Zealand

Abstract. In order to describe a journey, a story could be automatically generated from a group of digital photographs. Most of the existing methods focus on descriptions of specific content of a single image, such as image captioning, which lack of correlation between the images and the spatiotemporal relationships. To this end, in this paper, our goal is to propose a novel storytelling architecture based on computer vision. It makes use of visual object detection from digital images. Combining the changes in spatiotemporal domain and filling in the predetermined template, we automatically generate a text-based travel diary. In this project, compared with conventional image captioning, our aims are to effectively connect correlation between digital images and background information. The contributions of this paper are: (1) Innovative use of preset templates to generate travel diaries from photographs, associating content and context of the images as an event, (3) augmenting the images to expand the dataset, (4) shortening training time of deep learning models.

CCS CONCEPTS • Deep Learning • Object detection • Storytelling

ACM Reference Format:

Y. Zhu, W. Yan. Image-Based Storytelling Using Deep Learning. In ICCCV' 22: The 5th International Conference on Control and Computer Vision (ICCCV 2022), August 19-21, 2022. Xiamen, Fujian China.

1 INTRODUCTION

With the rapid development of Internet, a large amount of multimedia data is deposited. One of the most popular areas is how to let computers understand the information contained in digital images. An image from a digital camera is a large amount of color pixels in a huge matrix. Thus, we train computational models to understand and describe the images, known as image caption. Therefore, computing machines could describe the semantic information contained in each image. While storytelling is different from image captioning, a picture is only to record what we saw before a digital camera, but the storytelling is to write down how we feel and what we think, which could give us more options to share our memory. For those who love to travel, which can discover many views that have never been experienced before, such as delicious foods described in books, historical places of interest that have gone through thousands of years. The images recorded by a camera often have deep and long impact [1].

People will take the time to document a trip because the words behind the pictures are better 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, we can use photos to generate travel diaries more conveniently and quickly. We hope that this method will assist disables as well. In this paper, we demonstrate deep learning

models to automatically identify incoming images, extract visual features in the photos, and text-based generate travel diaries according to the timeline. The focus of this paper is how to use visual object detection in deep learning to detect special objects from the given photos and obtain story descriptions with emotions by using template-based languages. The target detection method will also be combined with Transformer. Throughout this paper, we contribute to: (1) Detect visual objects by using CNNs and Transformers, (2) analyze object detection algorithms with specific datasets, (3) conduct data augmentation (4) generate travel stories based on given photographs with visual objects, timelines, and emotional information.

The remaining part of this paper is organized as follows. Literature will be reviewed in Section 2. Our methodology is depicted in Section 3. Our result analysis is presented in Section 4. Our conclusion and future work are addressed in Section 5.

2 LITERATURE REVIEW

Pertaining to the travels, the photos were taken to record the specific time, scene, people and activities [1], what is recorded behind a photo can be written down as the context to describe a scene or a textual story automatically. By storytelling, we enjoy ourselves and others with a better understanding of culture, convey emotions and make the sightseeing trip much meaningful by using artificial intelligence.

In the summer of 1956, in an American conference at Dartmouth University, the concept of artificial intelligence was proposed, many years later, it was identified as the starting point of global artificial intelligence research [2, 3, 4]. Deep learning is based on a collection of machine learning algorithms [5].

The methods of deep learning are categorized as supervised [6] or unsupervised [7, 8, 9, 10]. LeCun et al. introduced the basic convolutional neural network (CNN), which includes convolutional layer, pooling layer and full connection layer [11, 12]. CNN has also become a reliable method in object detection, recognition and tracking [13, 14, 13]. In particular, Huang et al. proved that DenseNet greatly reduced the number of parameters [15], which is use of ReLU as the activation function and Dropout to reduce overfitting of the model [16].

CNN and RNN (i.e., recurrent neural network) are the two discriminative structures in deep learning [17]. RNNs are a class of networks that take sequential data as input, i.e., the previous input is related to the later input that are employed for natural language processing (NLP) related problems [18, 19]. Among them, Bi-directional RNN (Bi-RNN) [20, 21] and Long Short-Term Memory networks (LSTM) [22, 23] are the exemplar RNN models. The use of RNNs allows a connection between a word and its left and right adjacent words in semantics [24]. The LSTM net was developed to solve gradient vanishing problem, the LSTM actually replaces a neuron in the hidden layer of the RNN with a more complex structure called memory block [25, 26, 27].

Image captions like storytelling, which requires the analysis of images and the generation of textual descriptions based on the content of the given images. Describing an image in sentences is much more challenging than words, because it is often difficult to predict the relationships between different objects [28]. Generally speaking, existing image subtitle algorithms are grouped into three categories according to the ways of sentence generation: Template-based methods, transit-based methods and neural network-based methods [29]. The evaluation metrics include BLEU, METEOR, CIDEr, and SPICE [30]. Regarding 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 [31].

ResNet and DenseNet are very suitable for generating image captions with low complexity. The CNNs are applied to convert images to 1D vectors and extract visual features, the image features are mapped to the vector space of hidden state of LSTM, image feature vectors at each time-step are employed to calculate attention [32]. LSTM demonstrates the correct and reasonable combination of the importance of words, semantic objects and scenes [33]. 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 [34]. This method provides a much relevant and coherent NLP result. 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 [35].

Regarding visual object detection, the task results in the need to find and label a specific object within the image [33]. R-CNN for visual object detection was firstly proposed in 2014 [36, 37]. In 2015, a faster detector Fast R-CNN [38] is a method integrating with R-CNN and SPPNet [39] together. Followed by Faster R-CNN [40], a regional proposal network (RPN) was compared to Fast R-CNN.

Compared with the two-stage detection, a completely different one-stage detector YOLO was firstly proposed in 2015. Unlike previous proposal detection + validation methods, YOLO takes use of a single neural network to act on the entire image [41]. YOLO segments the given image into multiple regions and simultaneously predicts the probability that the bounding boxes of each region have been classified. If the center of an object is in this grid, the grid is responsible for predicting the object. Each bounding box needs to regress its own position and predict a confidence score. YOLO model supports the same input resolution as the training image during object detection. However, in subsequent updates, both YOLOv3 and YOLOv5 have created new milestones that have been greatly improved the computational speed and accuracy, the accuracy of identifying small objects has been improved.

3 METHODOLOGY

CNN is a type of feedforward neural networks as shown in Fig. 1. Different from ordinary fully connected feedforward 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 trainable weights and bias constants. Each neuron has its input, the output is employed for the final classification.

Figure 1: Conv model

In Fig. 1, pooling layers are employed after multiple convolutions. Pooling is also a popular operation in the convolution process. It is a down sampling method with the purpose of reducing feature dimensions. By

adding SPP after CNN, the full connection layer can also adapt to input images of different sizes. Visual features in a region are sampled quickly through pooling layer, the robustness of the model against translation or rotation is enhanced at the same time, because the output is calculated by values within the range. Max pooling was added for this project, with the goal of retaining texture information in the model.

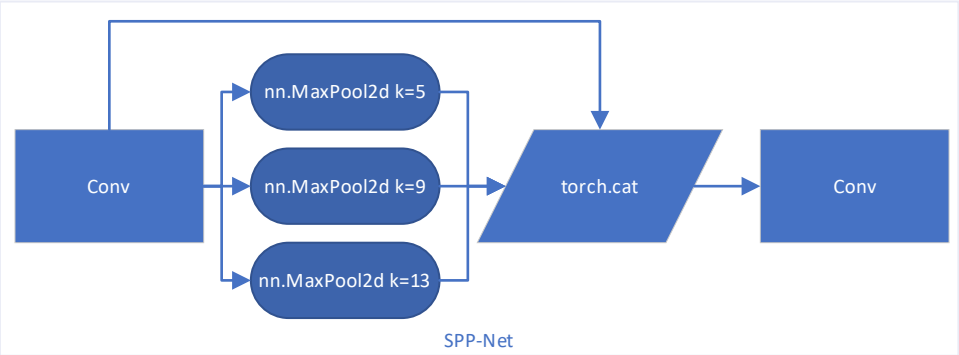


Figure 2: SPP-Net

Usually, after the convolution operation and before the full connection layer, we have the activation function. Throughout continuous learning, we solve the problems that the linear model cannot solve. As shown in the Fig. 1. the SiLU function shown in (1) has been taken into consideration in our experiment, SiLU outperformed than ReLU. 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.

$$SiLu\,(x) = x * sigmoid(beta * x) \tag{1}$$

In the process, Transformers have been explained for multiple times as an innovative way of target detection in image captioning. 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, through studying the relationships between vectors, we are able to determine which ones should be given more attention.

For the attention mode, it is essentially to convert the input image into a vector, then we translate it into a data structure Key and Value. By calculating the weight coefficient between Key K and value V, where V is weighted and summed to get the final attention. That is, Query and Key are applied to calculate the weight coefficients which take use of the function as shown in (2), t is time or sequence and d_k is hyperparameter.

$$Attention\,(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Multihead attention likes the number of X self-attention ensemble. Similar to convolution, convolution is use of convolution kernel, but Multihead attention takes advantage 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 current step depends on the hidden state of the previous step, which means it is a sequential procedure, if each calculation needs to wait after the previous calculation is completed. As an innovation, Transformer will be carried out in parallel, which could significantly improve the speed of training.

The hybrid structure enables Vision Transformer (ViT) to 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. Therefore, the Transformer module is used to replace the C3 layer in the last layer of the backbone. In other cases, an attempt to replace all C3 modules with transformers resulted to OOM error.

In visual object detection system, Deformable Parts Models (DPM) method was adopted to propose target region by sliding frame method, the classifier is adopted to realize recognition. In this paper, we take advantage of the one-stage detection method and directly output the scores and regression for anchors. The advantage 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 send 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 networks, reduces the number of model parameters and FLOPS, which not only ensures the inference speed and accuracy, but also alleviates the model size.

Bounding boxes (BBox) and confidence score of each grid cell are predicted. The confidence represents the probability that box contains a target. If there is no target, the confidence value is zero. In addition, the confidence value is as same as the ground truth intersection over union (IoU). Each grid cell is applied to predict the conditional probability. The probability represents the confidence that the grid contains a target. During the test, each box is multiplied by class probability and bounding box confidence to get a specific score. This score represents the probability of the class of object shown in the box.

In order to prevent overfitting of the model due to a small number of samples, the dropout method will be harnessed to randomly abandon a few of neural units, because dropout makes two neurons not always appear in the same subnetwork. In order to expand the diversity of samples, the samples will be randomly clipped and scaled. HSV, which stands for hue saturation value, is a color space based on the intuitive characteristics of colors. The data is expanded to simulate the color difference caused by digital cameras of different brands or mobile phones. These methods are employed to increase the parameters on the generalization ability of a data set, and improve robustness of the model.

Table 1: Pseudocode for generating textual story

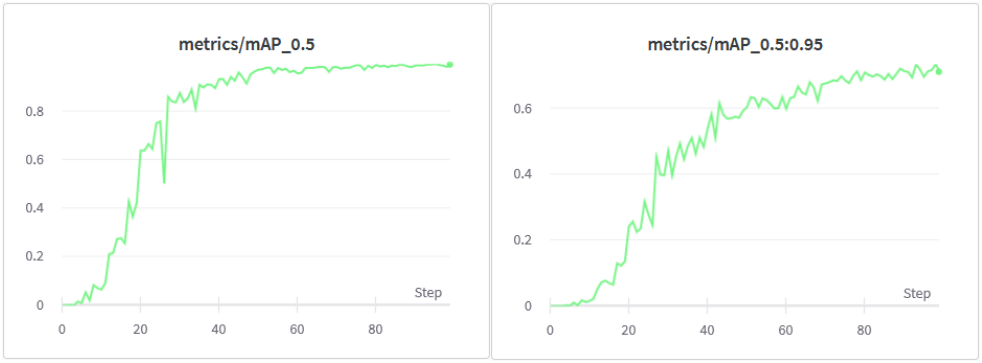
Input: Image training samples X, test samples Y **Output**: Class labels of Y
* Training Stage *
Learning the marked samples of all the photos; Using different algorithms to create suitable model; Prepare preset templates that mark up possible named entities;
* Testing Stage *
Detect object from input photos; Organize named entities according to the timeline; Fill the templates;

In order to present a realistic travel diary, we make use of a template to generate the story, which has the advantage of avoiding stiff sentences as shown in Table 1. 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 data is acquired and sorted out to fully describe the whole journey, 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 long 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. In addition, it will generate sentences based on relationships of positions according to the detected objects.

4 RESULT ANALYSIS

Pertaining to data collection, the famous tourist sites in Nanjing city 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 like human documentation. As a tourist, a lot of photos or short videos were taken during the journey of a day. Whilst making data annotation, the videos were parsed into pictures by extracting frames. The same target from different devices and 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. We collected 1,065 photos, for each picture, manual annotation is carried out to make the sample as accurate as possible.



(a) Result mAP@0.5 (b) Result mAP@0.5:0.95

Figure 3: Visual object detection using original model

Based on the training dataset, the result of object detection models is trained by using tourism photos which have the precision up to 0.9923 mAP that is considered to be able to effectively identify key object from images. It is very practical for extracting key objectives and generating tourism stories. The sample number was not large enough, after 80 iterations, the results tend to be stable. The accuracy and recall rates were close to 1.00, which means that our model did have high robustness. As we see from the results shown in Fig.3 related to mAP 0.50:0.95, the accuracy keeps rising, which shows that the boundary boxes predicted by the model still maintain the generalization ability.

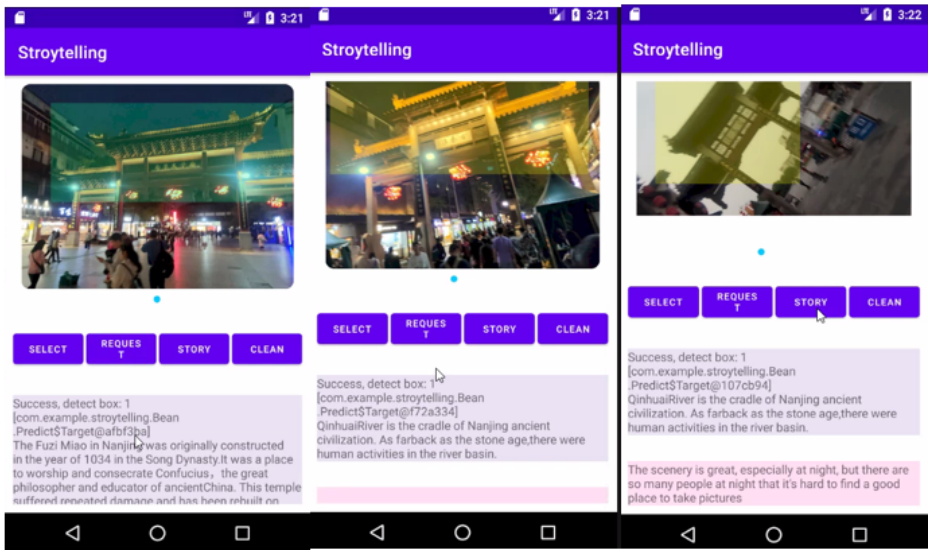


Figure 4: The screenshot of developed application and generated story

In this paper, the 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, human language is very subjective [31]. In this paper, our 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 timeline. 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 in Fig. 4 are Confucius Temple, Ancient Qinhuai and Nanjing Food stalls. Because we visited the whole scenic area, the story generated could summarize the whole journey and the background, or connect each photo based on the timeline such as when and where the photo was taken, whether the property of the object is a scenic spot or a restaurant, and takes use of appropriate connectives.

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, whilst training data and parameters remain unchanged, Transformer and attention models were added to the last layer of the backbone network as shown in Fig. 5.

Figure 5: The comparison between YOLOv5s and YOLOv5s added Transformer

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 color transformation scheme is maintained during the training process, the photos shot in the daytime are applied as the training set and the photos taken at night are used as the verification set to observe the generalization of the model shown as Fig. 6. 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 dataset was harnessed, 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.

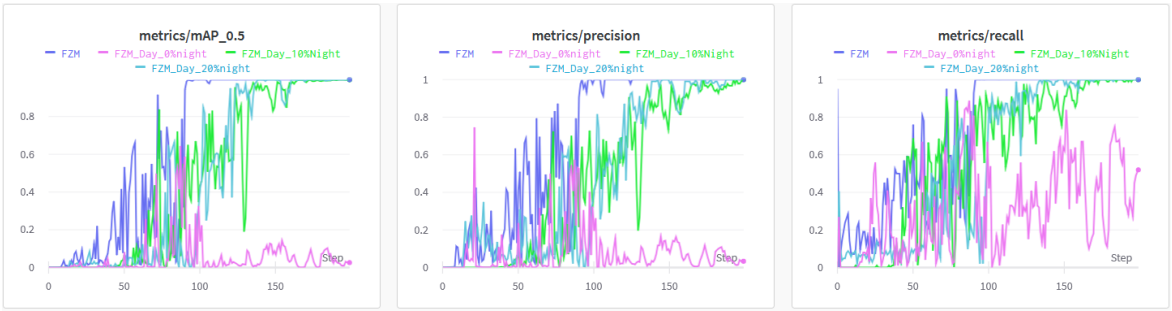


Figure 6: The comparison between adding different numbers of nighttime samples

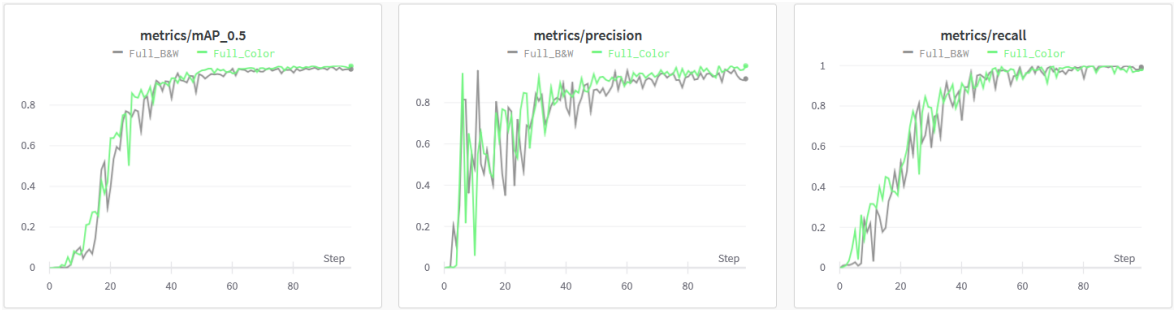


Figure 7: The comparison between colored images and black& white image

While switching to black and white colors, we see it had not 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, color filtering has not influence on performance and accuracy as shown in Fig. 7.

For most tourist photos, the object to be detected is usually not too small. Therefore, in previous experiments, the image resolution 640 x 640 was selected for the input image. In the case to accelerate model training and object recognition, smaller images are also an option. We resize image to 320 x 320 for training so as to explore the correlation between computational rate and speed.

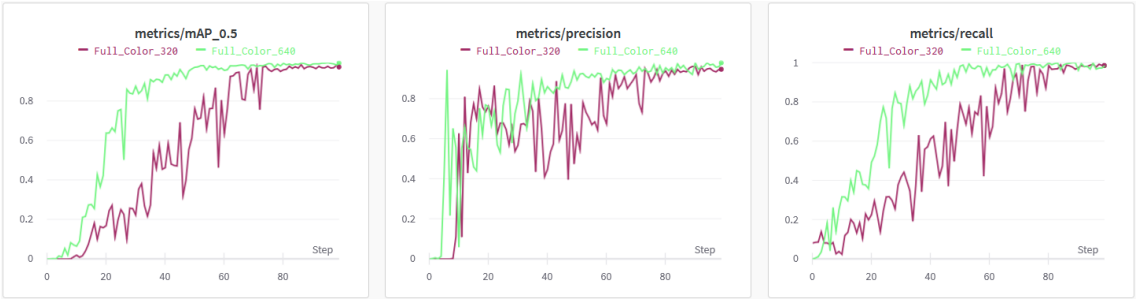


Figure 8: The comparison between colored image and black& white image

From the experimental results shown in Fig. 8, we see that when the image size is reduced from 640 x 640 to 320 x 320, Mean Average Precision (mAP) @ 0.50 was achieved as before, but with 20 epochs, only half of the time costed. However, by comparing with F1 curves, we found that there is almost no difference between the accuracy rates of different classifications with the 640 x 640 datasets. Therefore, a smaller model is considered for visual object detection in most scenarios.

5 CONCLUSION AND FUTURE WORK

The purpose of this paper 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. 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 is able to be applied to identify scenes and named entities in photos, and fill the named entities is a way of generating travel story.

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 completed, so the detection of partially overlap samples is not enough. The missing samples are not limited to partial overlap, 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. Therefore, optimization can be carried out for small object or overlap object to improve the mining of scene details [42]. In this paper, we extracted object contour after converting images into black and white. For example, we gave priority to extracting the contour of objects, but the effect was not good. This may be because the complex scene made it impossible to accurately determine the frame, therefore it need to find ROI from irrelevant context and misleading feature [43, 44]. More, filters and stickers will also have an impact on the photos, this part of the sample should also be added. Overall, more samples are added to improve the

robustness of the model. With the generated stories, there is also a need to consider the generality of templates in cases where they are dependent on them.

In addition, the way of storytelling combines text to language can also explore the world. In some travel scenes, we can expect that the image content extracted and converted into text or voice can be applied to some accessibility scenes, such as helping people with color blindness or blindness to better understand the road conditions [45]. 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 multidisciplinary integration to enable computers to understand images. Combining CNN, RNN, and Attention models together is a direction worth exploring, and LSTM can better fuse the context other than the image pixel content [46, 47, 48, 49, 50].

REFERENCES

[1] Marko Smilevski, Ilija Lalkovski, Gjorgji Madjarov. 2018. Stories for images-in- sequence by using visual and narrative components. In International Conference on Telecommunications, pages 148–159. Springer.

[2] Jack Copeland. 2015. Artificial Intelligence: A Philosophical Introduction. John Wiley & Sons

[3] Richard E Neapolitan and Xia Jiang. 2018. Artificial Intelligence: With An Introduction to Machine Learning. CRC Press

[4] Tom Mitchell. 1997. Machine Learning. McGraw Hill Burr Ridge

[5] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, Gulshan Kumar. 2020. A survey of deep learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering, 27(4):1071–1092

[6] Rich Caruana, Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In International Conference on Machine Learning, pages 161–168

[7] Geoffrey E Hinton, Terrence Joseph Sejnowski, et al. 1999. Unsupervised Learning: Foundations of Neural Computation. MIT Press.

[8] Geoffrey E Hinton. 2012. A practical guide to training restricted Boltzmann machines. In Neural networks: Tricks of the Trade, pages 599–619. Springer.

[9] Lu Zhang, Jianjun Tan, Dan Han, Hao Zhu. 2017. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. Drug Discovery Today, 22(11):1680–1685.

[10] Yujin Oh, Sangjoon Park, Jong Chul Ye. 2020. Deep learning COVID-19 features on CXR using limited training data sets. IEEE Transactions on Medical Imaging, 39(8):2688–2700.

[11] Yann LeCun, L´eon Bottou, Yoshua Bengio, Patrick Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278– 2324.

[12] Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. arXiv:1511.08458.

[13] Saad Albawi, Tareq Abed Mohammed, Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In International Conference on Engineering and Technology (ICET), pages 1–6. IEEE.

[14] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, Mohammed Bennamoun. 2018. A guide to convolutional neural networks for computer vision. Synthesis Lectures on Computer Vision, 8(1):1–207.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q Weinberger. 2017. Densely connected convolutional networks. In IEEE Conference on Computer Vision and Pattern Recognition, pages 4700–4708.

[16] Rahul Chauhan, Kamal Kumar Ghanshala, RC Joshi. 2018. Convolutional neural network (CNN) for image detection and recognition. In International Conference on Secure Cyber Computing and Communication (ICSCCC), pages 278–282. IEEE.

[17] Ju¨rgen Schmidhuber. 2015. Deep learning in neural networks: An overview. Neural Networks, 61:85–117.

[18] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, Noah A Smith. 2016. Recurrent neural network grammars. arXiv:1602.07776.

[19] Ian Goodfellow, Yoshua Bengio, Aaron Courville. 2016. Deep Learning. MIT Press.

[20] Manyu Dhyani, Rajiv Kumar. 2021. An intelligent chatbot using deep learning with bidirectional RNN and attention model. Materials Today, 34:817–824.

[21] Mike Schuster, Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681.

[22] Klaus Greff, Rupesh Srivastava, Jan Koutnik, Bas Steunebrink, and Jurgen Schmidhuber. 2016. LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10):2222–2232.

[23] Martin Sundermeyer, Ralf Schlu¨ter, Hermann Ney. 2012. LSTM neural networks for language modeling. In Annual Conference of The

International Speech Communi- cation Association.

[24] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In AAAI Conference on Artificial Intelligence.

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

[26] Leonard E Baum, George Sell. 1968. Growth transformations for functions on manifolds. Pacific Journal of Mathematics, 27(2):211–227.

[27] Mehdi Hosseinzadeh Aghdam. 2019. Context-aware recommender systems using hierarchical hidden Markov model. Physica A: Statistical Mechanics and Its Applications, 518:89–98.

[28] Abhinav Gupta, Larry S Davis. 2008. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In European Conference on Computer Vision, pages 16–29. Springer.

[29] Xu Jia, Efstratios Gavves, Basura Fernando, Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In IEEE International Conference on Computer Vision, pages 2407–2415.

[30] Xiaodong He, Li Deng. 2017. Deep learning for image-to-text generation: A technical overview. IEEE Signal Processing (Magazine), 34(6):109–116.

[31] Haoran Wang, Yue Zhang, Xiaosheng Yu. 2020. An overview of image caption generation methods. Computational Intelligence and Neuroscience, 2020.

[32] Sulabh Katiyar, Samir Kumar Borgohain. 2021. Comparative evaluation of CNN architectures for image caption generation. arXiv:2102.11506.

[33] Xintao Ding, Yonglong Luo, Qingying Yu, Qingde Li, Yongqiang Cheng, Robert Munnoch, Dongfei Xue, Guorong Cai. 2017. Indoor object recognition using pre-trained convolutional neural network. In International Conference on Automation and Computing (ICAC), pages 1–6.

[34] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, Qi Tian 2017. Image caption with global-local attention. In AAAI Conference on Artificial Intelligence.

[35] Ali Furkan Biten, Lluis Gomez, Mar¸cal Rusinol, Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12466–12475.

[36] Xiaoxu Liu, Wei Qi Yan. 2020. Vehicle-related scene segmentation using CapsNets. In International Conference on Image and Vision Computing New Zealand.

[37] Ross Girshick. 2015. Fast R-CNN. In IEEE International Conference on Computer Vision, pages 1440–1448.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9):1904–1916.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28:91–99.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788

[41] Ying Liu, Luyao Geng, Weidong Zhang, Yanchao Gong, and Zhijie Xu. 2021. Survey of video based small target detection, Journal of Image and Graphics, Vol. 9, No. 4, pp. 122-134.

[42] Rafflesia Khan, Tarannum Fariha Raisa, and Rameswar Debnath. 2018. An efficient contour based fine-grained algorithm for multicategory object detection, Journal of Image and Graphics, Vol. 6, No. 2, pp. 127-136.

[43] Teerapat Chaloeivoot and Suebskul Phiphobmongkol. 2016. Building detection from terrestrial images, Journal of Image and Graphics, Vol. 4, No. 1, pp. 46-50.

[44] Yanzhao Zhu and Wei Qi Yan. 2022. Traffic Sign Recognition Based on Deep Learning. Multimedia Tools and Applications, Springer.

[45] Ryo Hasegawa, Yutaro Iwamoto, and Yen-Wei Chen. 2020. Robust Japanese road sign detection and recognition in complex scenes using convolutional neural networks. Journal of Image and Graphics, Vol. 8, No. 3, pp. 59-66.

[46] Yulin Zhu. 2022. Image-Based Storytelling for Tourist Using Deep Learning. Research Project Report, Auckland University of Technology, New Zealand.

[47] Wei Qi Yan. 2021. Computational Methods for Deep Learning: Theoretic, Practice and Applications. Springer.

[48] Wei Qi Yan. 2019. Introduction to Intelligent Surveillance Surveillance Data Capture, Transmission, and Analytics. Springer.

[49] Chen Pan, Jianfeng Liu, Wei Qi Yan. 2021. Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.

[50] Chen Pan, Wei Qi Yan. 2020. Object detection based on saturation of visual perception. Multimedia Tools and Applications 79 (27-28), 19925-19944.