Problem Chosen
C

2021
MCM/ICM
Summary Sheet

Team Control Number
2113265

Bumblebee prediction based on LSTM network and VGG19 pre-trained convolutional neural network

Summary

As an invasive species, Vespa mandarine appears in Washington State. Enthusiastic citizens provide sighting reports for the government's follow-up investigations while the accuracy of sighting reports is poor. In order to efficiently and accurately use these data and predict the future distribution of V.M, we combine models such as LSTM network,VGG19 pre-trained convolutional neural network and so on. The solutions will contribute to prevent the harmful effects of invading wasps on the environment and economic development in advance, which is of practical significance.

For **problem one**, It is a predictive problem. We firstly expand the sample data by interpolation. And then we propose a model reflecting the relationship between the spatial position and temporal change of wasps based on the LSTM algorithm. Experiments show that our model has 62% accuracy. Finally,we simulate the potential distribution of wasps in the future based on this model. It provides a theoretical basis for wasp control.

For **problem two**, we construct a wasp classification model. Due to the large difference in the number of positive and negative samples, we use various methods to expand the data, such as translation, rotation and adding noise. Meanwhile, the non-wasp data is down-sampled. Next,on a small sample of wasp data,we conduct a classification model of wasp image based on the pre-trained VGG convolutional neural network by transfer learning. Our model can achieve an accuracy of more than 92% in the experiment so that it can effectively detect wasps with small errors.

For **problem three**, we start with a new string consisting of 50 most frequent words in the Negative ID sample and call the new one the Negative dictionary string.Second. TFIDF is used to extract the text word vector from the target report. We calculate the similarity between the target text and Negative dictionary string and compare similarity with the threshold. Finally, combined with the image information results in Model 2, three response levels are constructed to achieve a more objective response. We test Unprocessed and unverified samples in the data.Experiments show that our method can effectively divide response levels and greatly save time of manual judgment.

For **problem four**, we consider that deep learning model training often requires a large amount of data, so we combine the temperature situation and the life habits of wasps to make statistics on the report data generated every month. On the premise of ensuring effective data volume, we believe that the model should be updated as soon as possible.After analyzing the data, we find that the reports are mainly produced from April to October, especially the peak in August to September. Therefore, we conclude that updates should be made between August and September every year, and the frequency of updates should be once a year.

For **problem five**, we think of it as a concrete application of the previous model. Once there are no more primary and secondary responses, the pest is eradicated.

Keywords: Long short term memory(LSTM) network; VGG19 pre-trained CNN;Word cloud

Contents

1 Introduction 2

1.1 Problem Background 2

1.2 Restatement of the Problem 3

1.3 Literature Review 3

2 Preparation of the Models 3

2.1 Assumptions and Justifications 3

2.2 Notions and Symbol Description 4

3 Model Establishment and Solution of Problems 5

3.1 Data overview 5

3.2 Problem 1 5

3.2.1 Data preprocessing 5

3.2.2 Problem Modeling 7

3.2.3 Model solving 9

3.2.4 Conclusion 9

3.3 Problem 2 9

3.3.1 Data preprocessing 9

3.3.2 Problem Modeling 11

3.3.3 Model solving 12

3.3.4 Conclusion 13

3.4 Problem 3 13

3.4.1 Data preprocessing 13

3.4.2 Problem Modeling 15

3.4.3 Model solving 16

3.4.4 Conclusion 17

3.5 Problem 4 17

3.5.1 Data preprocessing 17

3.5.2 Problem Analysis and Modeling 18

3.5.3 Conclusion 18

3.6 Problem 5 18

4 Strengths and Weaknesses 19

4.1 Strengths: 19

4.2 Weaknesses: 19

5 Memorandum 20

Appendices 22

Appendix A 22

Appendix B 23

1 Introduction

1.1 Problem Background

The destruction of biodiversity caused by the increasing species invasion has been the focus of recent years. In addition to being one of the greatest threats to biodiversity and ecosystems, biological invasions have also caused enormous losses to the global economy. Invasive species have adverse effects on neocolonial regions, such as impoverishing native species assemblages in terrestrial and aquatic habitats, and decreasing key insects that supply biological control and pollution services. Due to their impact on local honey bee colonies, and possible negative human health effects from stings and their allergenic effects. The increase in invasive species and their impact on biodiversity and ecosystems raise a number of management and control issues.

A typical example is the emergence of Asian giant wasps on Vancouver Island in Canada and Washington State in the US, severely injuring local species in succession. Bumblebees of the genus Vespa are known to have serious negative effects on honey bee colonies, hunting foragers while attacking hives. These hybrid bumblebees live in social colonies established by female bees and can produce colonies in areas of human modification such as agricultural or urban landscapes. Because its nest is usually built underground, it is very hidden and difficult to find. At the same time, its appearance and life cycle similar to those of local bees interfere with human judgment.

It is recognized to prevent the establishment and further spread of invasive species. In order to avoid harmful economic and environmental consequences, Washington State plans to further investigate and study the specific scope of the invasion through the hotline and the information posted on the website by enthusiastic citizens, and then use limited resources to guide future eradication attempts. Figure 1 shows a simplified detection location map of Asian bumblebee. The result is a large number of false sights and only a few confirmed sights, which is not optimistic.

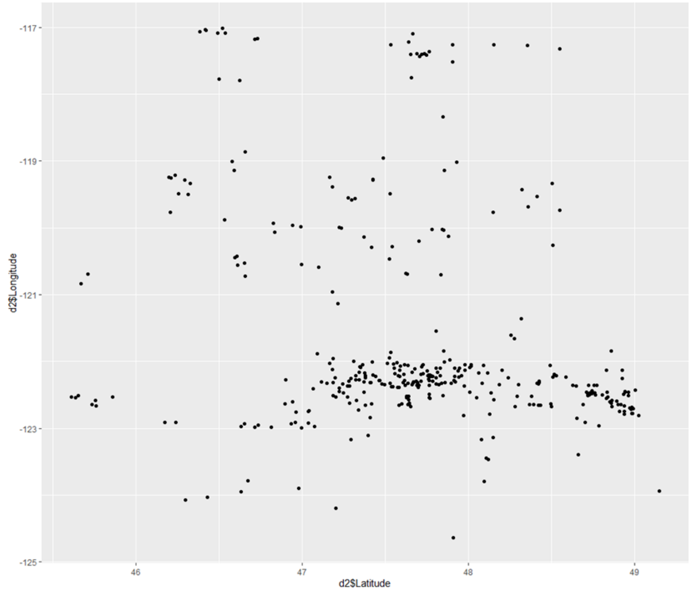


Figure 1: Map depicting Asian Giant Hornet Detections, as well as Hornet Watch and Public Sitting locations.After the simplification

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

Problem 1 Reliably predict the potential spread of the Hornet in North America in the future, and point out the accuracy of the forecast. Namelysolve and discuss whether it is possible to predict the spread of this pest over time.

Problem 2 Only use the data set and image files provided in the attachment to generate a predictive model. This model can point out the possibility that eyewitnesses will mistake the bee they see for the Asian Hornet, which is the possibility of anticipating misclassification.

Problem 3 Apply the established model to analyze the information provided by the witnesses, and explain that the events selected by the model that should be given priority investigation privileges are most likely to be the correct working principles of the witnesses.

Problem 4 When there are more reports of sightings, determine how to update the model to accommodate the analysis of the latest report, and determine the frequency of updating the model.

Problem 5 Through the analysis of the model created, it is indicated out under what situation and evidence that the pests were eradicated by Washington State.

1.3 Literature Review

Prior to this, the Asian Hornet was mainly anticipated based on objective nat-ural factors.

First from the point of view of the abiotic part , many scholars have analyzed and studied different ecosystem factors related to the Asian Hornet. Norderud , Powell & Peterson(2021) estimate for Asian Hornet queens to overwinter and colonize nests, density and distribution of apiaries based on climate and habitat suitability. In addition, by analyzing the organ-isms related to the Asian Hornet ,Gonzalez et al. (2021) thought the North American Asian bumblebee habitat is expected to roughly coincide with the highest honey production areas. This provides prey for the Asian Hornet. Futhermore, many people have proposed corresponding prediction methods and models. For example, based on the environmental factors,Jin and Li(2015) provided forecasting methods, such as experiential forecasting, ex-perimental forecasting and statistical forecasting. According to Morgane et al.(2018), The SDMS predicts the potential range of invasive species, using the Asian bumblebee as an example. The accuracy needs to be improved. However, when data is obtained, it is likely that its image will need to be clas-sified with other images of wasps that may be confusing. Kwon and Lee 2020gave the recognition score and classification accuracy of the fused wasp image based on the Deep Convolutional Neural Networks (DCNN) AlexNet algorithm.

2 Preparation of the Models

2.1 Assumptions and Justifications

To simplify our model and eliminate the complexity, we make the following main assumptions in this literature. All assumptions will be re-emphasized once they are used in the construction of our model.

1. **We assume that disregarding the influence of specific external factors such as climate.**

   Only the combined effect of all influences is recognized in the forecast.

2. **We assume that the past of things will continue into the future as well.**

   First, the development of things generally follows objective laws.The factors that influence the distribution of Asian bumblebees will remain almost the same as the past ones, so that their distribution can continue to follow past patterns.In other words, bumblebee distribution will not change in sudden jumps in the future, but in relatively small steps.Second, past and present phenomena may indicate trends in the development and change of present and future activities.

3. **We assume that the queen bee is inactive in winter.**

   Like other wasps, Asian bumblebees have a biological clock of one year. During the cold winter, the new queen leaves the nest and spends the winter in the soil. The only survivor waiting for spring is the wintering queen bee. In the spring, the queen bee is active and builds nests. This indicates that the queen is inactive during the winter.

4. **We assume that the flying distance of the queen bee during breeding is 30km.**

   We assume that the past of things will continue into the future as well.

2.2 Notions and Symbol Description

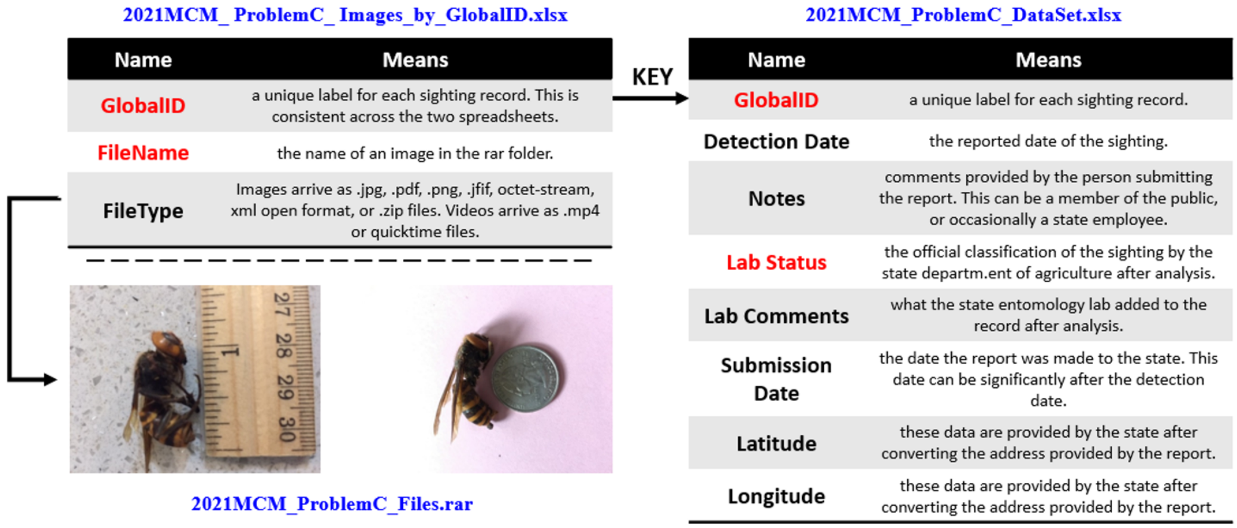
We will define the following variables here as they are used throughout our paper.

Symbols	Definitions
L^N	the newly generated data
L^B	the latitude and longitude of the adjacent start date
L^E	the latitude and longitude of the adjacent end date
h	the date of the difference between L^B and L^E
σ	the activation function, generally a Sigmoid function
W_k	the weight parameter of the hidden layer nodes, $k = f, i, c, o$
U_k	the weight coefficient of the input data, $k = f, i, c, o$
V_k	the weight coefficient of the LSTM neural network memory cell, $k = f, i, c, o$
c_{t-1}	the output of the LSTM neural network memory cell at the previous moment
b_k	the bias term, $k = f, i, c, o$
n	the number of samples
$\hat{y}_i$	the predicted value
y_i	the true value
TP	number of positive sarmples predicted by the model to be positive
FP	number of negative samples predicted by the model to be positive
FN	number of positive samples predicted by the model to be negative
TN	number of negative samples predicted by the model to be negative

3 Model Establishment and Solution of Problems

3.1 Data overview

We interpret and classify the names and meanings of the data in the two given datasets. Consequently, we figure out the key data and connection between them. The specific content is listed in the figure.



2021MCM_ ProblemC_ Images_by_GlobalID.xlsx

Name	Means
GlobalID	a unique label for each sighting record. This is consistent across the two spreadsheets.
FileName	the name of an image in the rar folder.
FileType	Images arrive as .jpg, .pdf, .png, .jfif, octet-stream, xml open format, or .zip files. Videos arrive as .mp4 or quicktime files.

2021MCM_ProblemC_Files.rar

2021MCM_ProblemC_DataSet.xlsx

Name	Means
GlobalID	a unique label for each sighting record.
Detection Date	the reported date of the sighting.
Notes	comments provided by the person submitting the report. This can be a member of the public, or occasionally a state employee.
Lab Status	the official classification of the sighting by the state departm.ent of agriculture after analysis.
Lab Comments	what the state entomology lab added to the record after analysis.
Submission Date	the date the report was made to the state. This date can be significantly after the detection date.
Latitude	these data are provided by the state after converting the address provided by the report.
Longitude	these data are provided by the state after converting the address provided by the report.

Figure 2: The relationship each data set, the red letter represents the key data

3.2 Problem 1

To address and discuss whether the spread of this pest over time can be predicted and to what degree of accuracy, we decided to use multivariate time series analysis with the date of reported sightings as time and the latitude and longitude of the sighting site as bivariate variables.

3.2.1 Data preprocessing

We analyze the data given in the files and count the number of classified samples of the Lab Status item in 2021 MCM_ProblemC_DataSet.xlsx, as shown in the following table:

Table 1: Classification Sample Statistics of Lab State

Lab Status	Means	Numbers
Positive ID	is confirmed as an Asian Giant Hornet.	14
Negative ID	is excluded	2070
Unprocessed	not yet been classified.	15
Unverified	no determination was made due to lack of information.	2342

The information provided by Positive IDs is worth concerning.

Considering that the flying distance of the bumblebee is 30km during breeding, we draw their trajectory. From figure(a), we can find that the Global Id of (FC6E894B-F6DF-4FDC-853A-D7372D253988) and (124B9BFA-7F7B-4B8E-8A56-42E067F0F72) are far greater than the effective flight distance of the Hornet. Therefore, we regard them as outlier elimination, and draw revised trajectory diagram of the bumblebee (b).

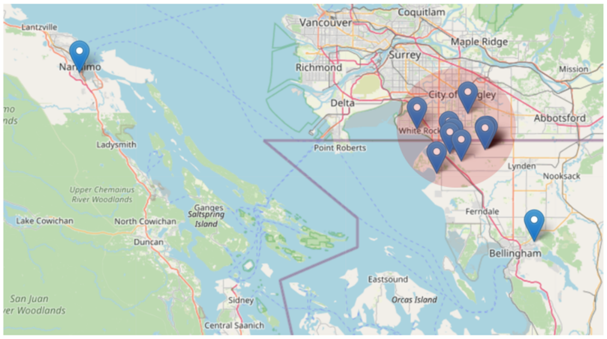


Figure 4: (a) represents the effective range of the bumblebee

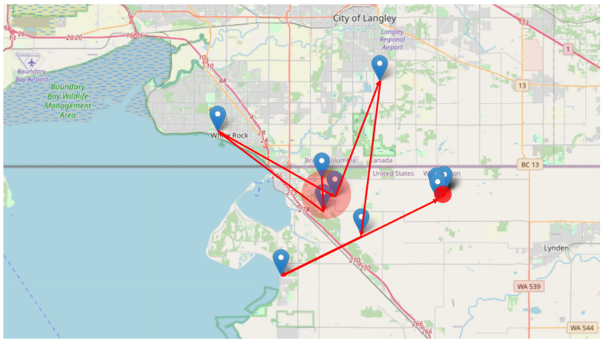


Figure 5: (b) represents the trajectory of the bumblebee

Next, we extend the records of Positive IDs.After analyzing the key attribute of Positive IDs and Detection IDs, we find that the interval between detection times is irregular.It is known from problem literature that the queen bee is inactive in winter. Therefore, we do not supplement the data in winter. For the rest of the data, we fill in the mean interpolation at 15-day intervals.

The filling formula is as follows:

$$L^N = L^B + \frac{L^E - L^B}{\frac{h}{15}} \tag{1}$$

Part of the expanded data is shown in the following table:

Table 2: Some examples of expanded data

Detection Date	Latitude	Longitude	Lab Status
2019/10/15	48.98292	-122.702	Positive
2019/12/1	49.00341	-122.75	Positive

At the same time, in order to analyze the relationship between a bumblebee's moving position and time, we construct a time series data set. And then, we use the former three sets of data to predict the latter one The purpose of constructing the data set in this way is to fully consider the feature correlation between the spatial position and time of the bumblebee.

3.2.2 Problem Modeling

In response to the first question, we constructed a sequence data set of the bumblebee's moving position and time, and at the same time built an LSTM network to capture the relationship between the bumblebee's spatial position and time, and then realized the accurate prediction of the bumblebee's spatial position change.

The structure of our model is shown in the following diagram:

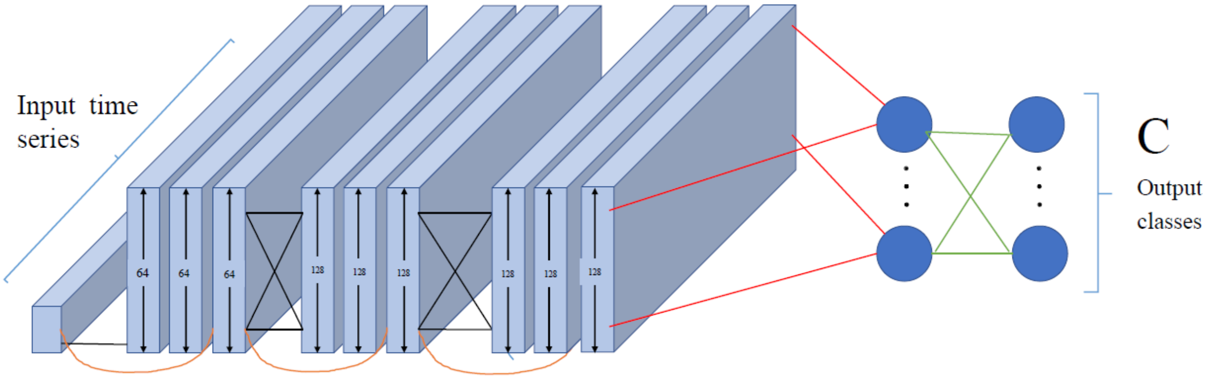


Figure 5: Residual network's architecture. Black connections correspond to convolutions, orange to residual, red to average pooling and green are fully-connected

The long short term memory (LSTM) neural network, first proposed in 1997, is a huge improvement for the recurrent neural network (RNN).

As shown in the figure 5, through selective memory of historical information, LSTM neural network solves the problem of RNN gradient explosion and gradient disappearance. LSTM neural network adds three important gate control functions to the memory unit module of the hidden layer. The three different gate control functions correspond to three different gate control functions, namely input gate, forget gate and output gate. The LSTM neural network controls the storage and transmission of information through three gate functions of the hidden layer memory module.

(1) The main role of the forgetting gate (f_t) is to control the amount of data of the input information and to control the time span of the data consistently by discarding part of the historical information, which is achieved by the Sigmoid activation function and takes values between (0,1). Figure 6 shows the schematic diagram of the forgetting gate structure

The forgetting gate controls the superposition of the hidden layer h_{t-1} output and the input layer x_t input by the sigmoid function, and the output f_t of the forgetting gate is expressed as

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + V_f c_{t-1} + b_f) \tag{2}$$

(2) The main role of input gate ((t_i)) is to update the information of the current moment and input the output of the hidden layer of the previous moment into the current moment, so as to achieve the transfer of timing information.Figure 7 shows the schematic structure of the input gate.

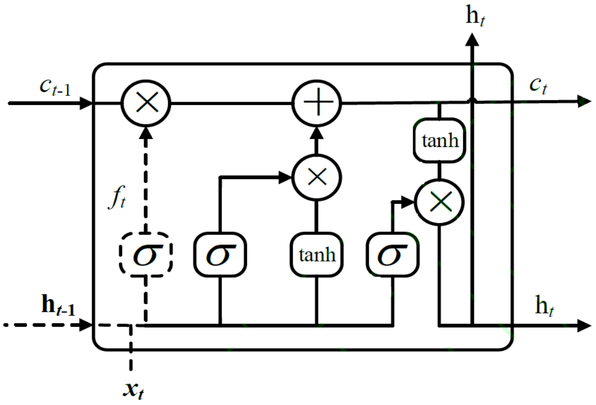


Figure 7: The sketch of input gate of LSTM

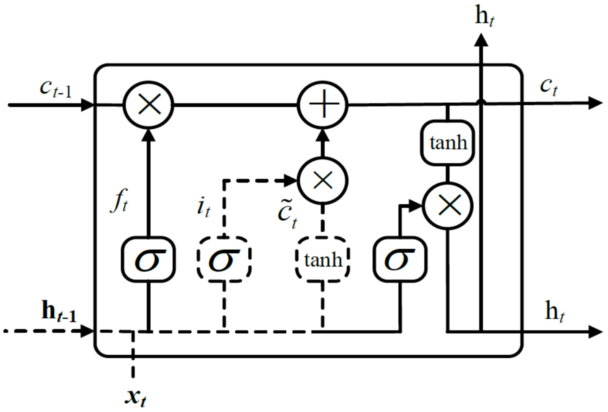


Figure 8: The sketch of input gate of LSTM

The process of superimposing the output of the hidden layer at the previous moment with the current input is

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + V_i c_{t-1} + b_i) \tag{3}$$

According to the joint action of the forgetting gate and the input gate, the hidden layer memory cell realizes the information update and outputs the memory cell c_t , and the hidden layer memory cell is calculated as

$$c_t = i_t \tanh(U_c x_t + W_c h_{t-1} + b_c) + f_t c_{t-1} \tag{4}$$

(3) The output gate (o_t) controls the output information and returns the timing information of the hidden layer through the output gate before the memory cell outputs the information, thus ensuring that the information of the memory cell is updated. The structure of the output gate is shown in Figure 8.

The output gate is calculated as

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + V_o c_t + b_o) \tag{5}$$

The method to return the output h_t of the hidden layer is calculated as $h_t = o_t \tanh(c_t)$

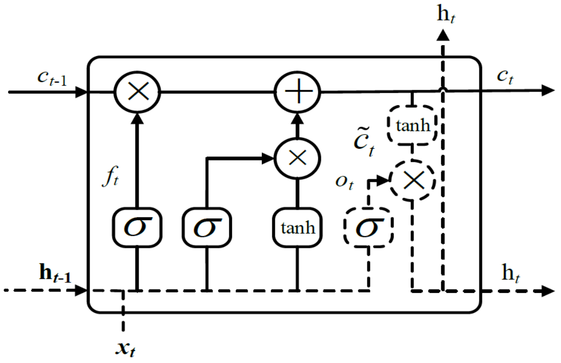


Figure 8: The sketch of input gate of LSTM

3.2.3 Model solving

We use the expanded time series data set about hornet to train the LSTM model. We use keras library in python to build LSTM to predict the spatial position of the bumblebee. See the appendix for the specific code. First, we set the number of single-layer LSTM units to 32. Then we set the dropout layer to 0.15 to randomly discard 15% of the parameters. The batch size for model training is 4 and the epoch is 15. At the same time, Mean Square Error (MSE) is used as the loss function, and Adam is the optimizer.

The calculation formula of MSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{6}$$

In the end,after 15 rounds of calculation, the final MSE value of the training set is calculated to be 0.1531. The accuracy rate reached 66.7%. This means that our model can predict the spatial position of the wasp.

3.2.4 Conclusion

After using the LSTM model, the prediction accuracy rate reaches more than 60%. This model can be used to predict the spatial position of the wasp.

3.3 Problem 2

3.3.1 Data preprocessing

We first identify the Global ID as the key factor in the process of connecting simultaneous image and label data tables, and then perform an overview analysis of the data.The key data items are shown in the table below:

Table 3: Data overview table

Name	Images Numbers	GlobalID Numbers
PositiveId	14	11
NegativeId	3194	2126
Unprocessed	11	-
Unverified	86	-

It can be seen from the files that the number of image samples of non-hornets is much more than that of hornets. In deep learning, for avoiding overfitting, we should input sufficient data. So we first increase the number of bumblebee image samples.

Considering that the bumblebee image contains redundant data, we cut the bumblebee image to ensure that the data redundancy after data enhancement is less in advance.The examples of image cropping processing are shown below:

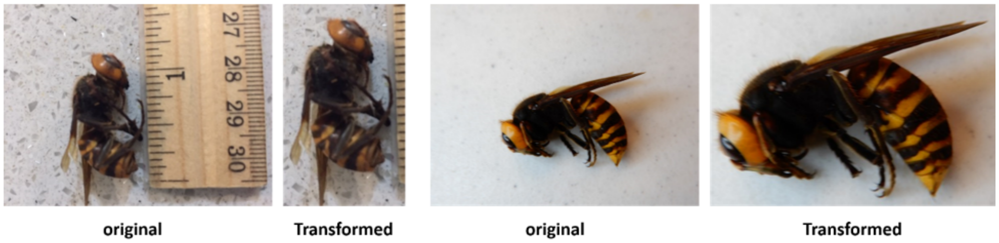


Figure 9: cropping example

We enhance the accuracy of the data by using the following methods:

- **Rotation:** Randomly rotate the image to a certain angle; change the orientation of the image content;
- **Flip:** Flip the image along the horizontal or vertical direction;
- **Zoom:** Zoom in or zoom out the image according to a certain ratio;
- **Shift:** The image is translated in a certain way on the image plane;
- **Scale:** The image is enlarged or reduced according to the specified scale factor. Change the size or blur degree of the image content;
- **Contrast transformation:** In the HSV color space of the image, the saturation S and V brightness components are changed, and the hue H remains unchanged. Exponential operations are performed on the S and V components of each pixel to increase the illumination change.
- **Noise:** Randomly perturb each pixel RGB of the image, we use Gaussian noise.

We transform each image 40 times, and the transformed image output is shown in the following figure:

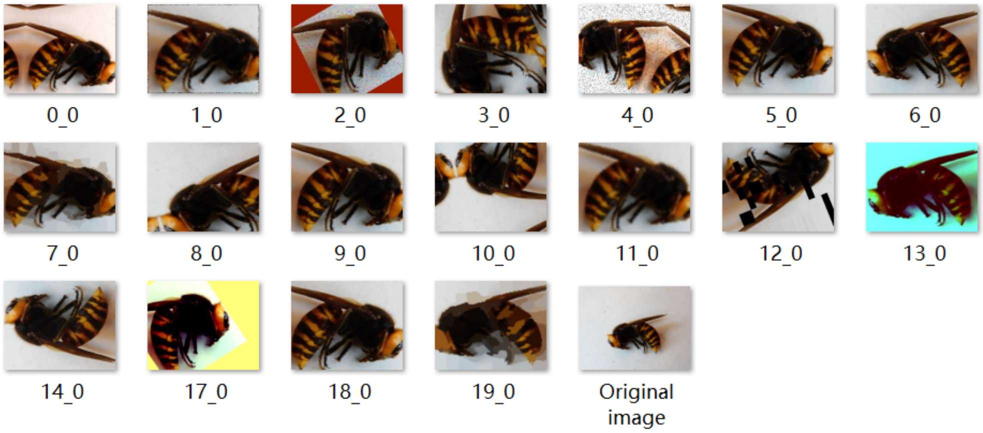


Figure 10: Transformation result output

At the same time, too many non-bumblebee samples are easy to cause data redundancy.This redundancy can also happen because the images provided by the same Global ID are too similar.Therefore, if there are multiple images in a global ID, we will randomly select one of the images for subsampling to ensure the balance of positive and negative samples.

3.3.2 Problem Modeling

In the second question, considering that samples is few in number, we combine the knowledge of transfer learning and use the VGG19 pre-trained convolutional neural network to perform transfer learning. The model structure is shown in the figure below:

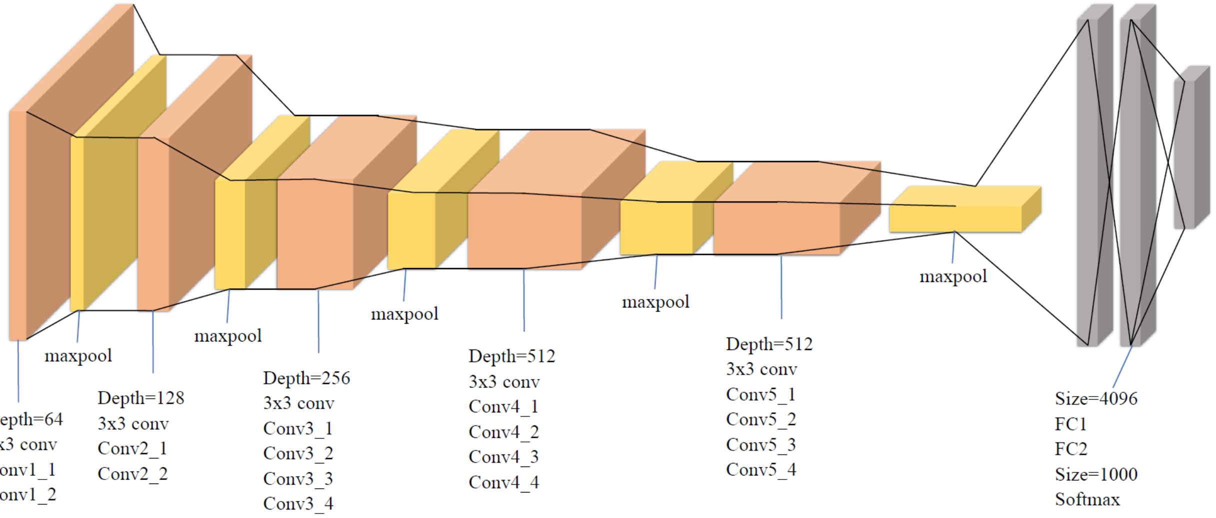


Figure 11: VGG-19 Model Architecture

Transfer learning is a machine learning method that can apply knowledge learned from other fields to the required research field. When a neural network is trained, it is generally considered that the training set and the test set are the same distribution, and thousands of samples are needed to train and optimize the model. Therefore, even if the samples are sufficient, processing the sample data is a costly process; and the transfer learning method can well make up for the problems caused by insufficient data or excessive training time.

The VGG16 convolutional neural network is a process of extracting and continuously refining images, and it has a deep network structure.At first, the convolutional layer only extracts the features of the image, and only when the network layer is deeper can it deal with specific tasks. Therefore, in the migration learning process, the model that has been pre-trained in the source domain can be used, the weight of the lower layer is retained, and only the upper layer is retrained and related parameters are fine-tuned. Because of the different characteristics of the source domain data distribution, it is necessary to use the target domain dataset to fine-tune the model, remove the original top layer and add a new output layer, and add the softmax function to classify the new problem. Figure 12 shows a process of migrating the model. By migrating the weights trained in the source domain (natural image) in the model, and then using the target domain (task data) for fine-tuning, it greatly simplifies the model in the new domain.

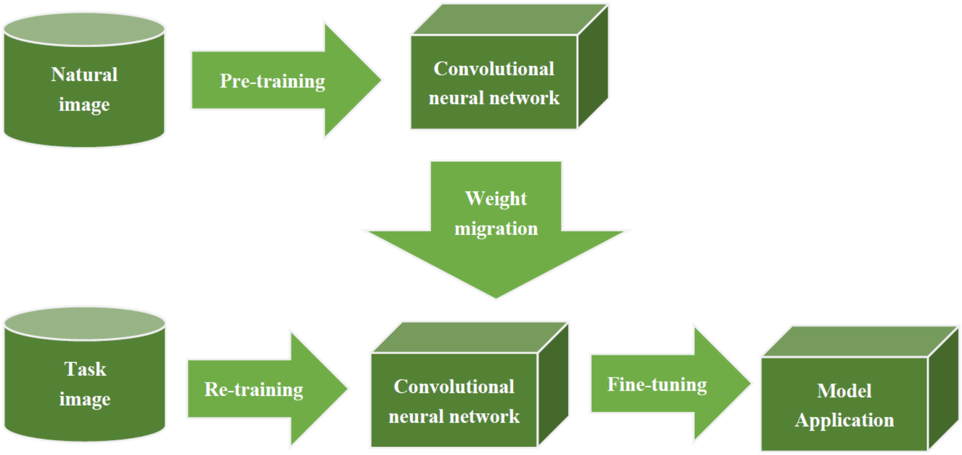


Figure 12: Model transfer of CNN

3.3.3 Model solving

The new data set after data enhancement is used to train the convolutional neural network for image recognition of bumblebee. We use the VGG19 model pre-trained on imagenet. However our training strategy is fine-tune. Specifically, we freeze the parameters of the first fifteen layers of networks, and then retrain the models after fifteen layers. It ensures that a better effect can be achieved on small data sets.

We divide the data set into training set and test set according to 8:2. When the batch size is set to 32 and the learning rate is 1e-4, 20 epochs are trained. At the same time, we use the binary cross entropy loss function, the formula is as follows:

$$L_{logistic}(\hat{y}-y)=-y\log\hat{y}-(1-y)\log(1-\hat{y}),\ y\in\{0,1\} \tag{7}$$

$\hat{y}=\sigma(\bar{y}),\ \ \sigma(\bar{y})=\frac{1}{1+e^{-\bar{y}}},\ \ \frac{\partial\sigma}{\partial\bar{y}}=\sigma(1-\sigma),\ \ \bar{y}=w\cdot x+b=\sum_j w_j x_j+b,\ \ \frac{\partial\bar{y}}{\partial w_j}=x_j$

During the model training process, we draw the test set , training set loss and accuracy change curve. They are in the figure 13.

It can be seen that after the 8th epoch, the accuracy of the model on the test set basically does not change. Meanwhile, the model reaches a state of convergence. At this time, the accuracy on the test set is 92.7%. This strongly proves the effectiveness of our data enhancement strategy and the precise identification of bumblebees.

Besideswe also use multiple evaluation indicators such as F1 score precall and recall to test the error rate of the model.

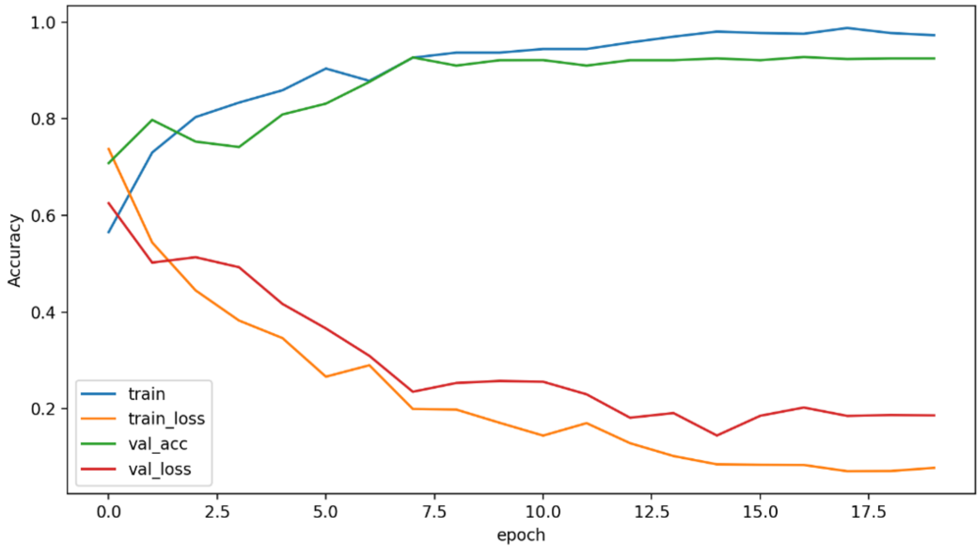


Figure 13: Change curve of training set loss and accuracy

The relevant calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precall = \frac{TP}{TP + FP}, \; Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precall \times Recall}{Precall + Recall} \tag{10}$$

The final model's performance test results for each indicator on the test set were accuracy rate 0.927, precall: 0.893, recall: 0, 915, and F1 score: 0.903.

3.3.4 Conclusion

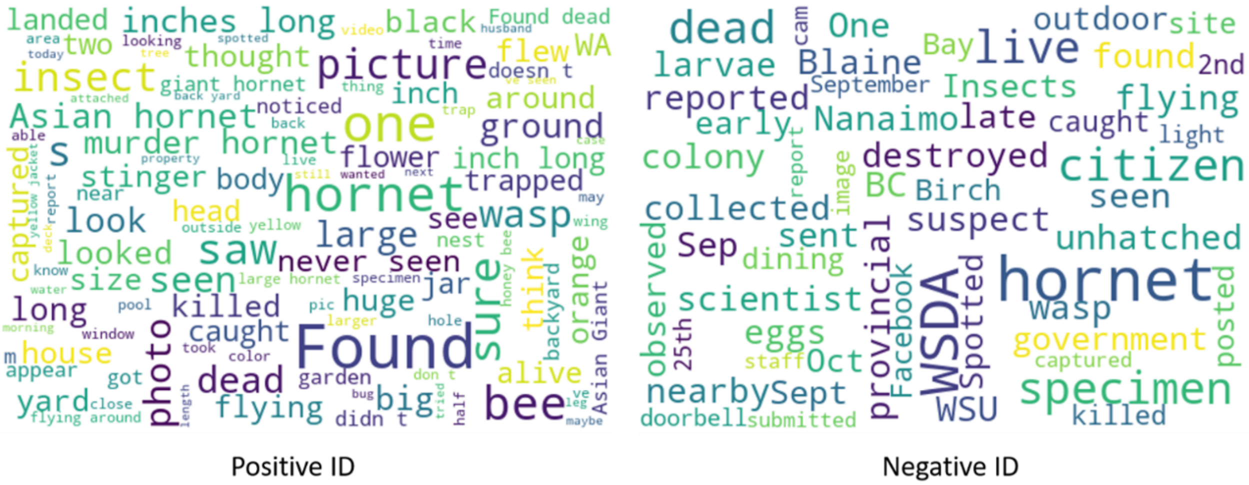
From the results, the model can be used to determine whether the bee species witnessed by the eyewitness is the target bee species.

3.4 Problem 3

3.4.1 Data preprocessing

In the modeling of question three, we comprehensively considered the image information and text information of each event to establish an event quick response model to detect whether the unprocessed and unprocessed data contains bumblebee-related information.

We first filter the data whose Lab Status is negative and positive, and then use the jieba module in python for word segmentation of the text in their note attributes, and use wordcloud to draw their respective word clouds for data overview. The cloud is shown below:



Positive ID

Negative ID

Figure 14: High-frequency word cloud image

Then we calculated the 50 most frequent words in negative, and the word frequency distribution diagram is as follows:

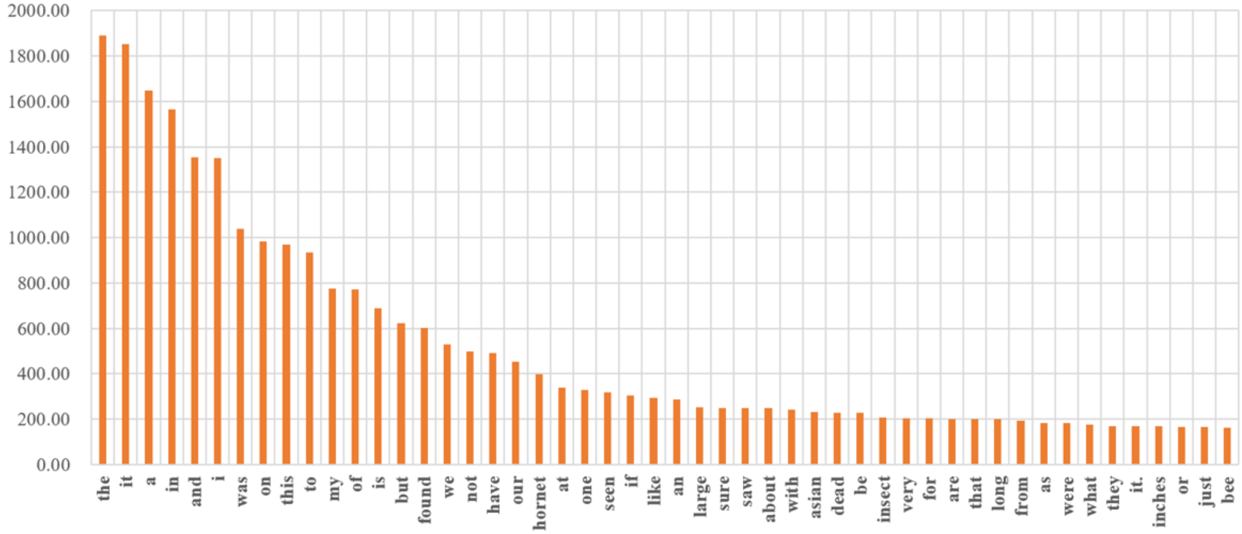


Figure 15: Word frequency distribution map

We compose a new string of the fifty most frequently occurring words as a negative dictionary string, and compare the similarity of the note text in other records with it during model building. The samples with high similarity are considered to be negative samples, and the samples with low similarity are considered to be positive samples. The specific model design and threshold selection are explained below.

3.4.2 Problem Modeling

Under the premise of limited manpower and material resources, in order to quickly judge the right or wrong of eyewitness reports, we built a time quick response model as shown in following figure:

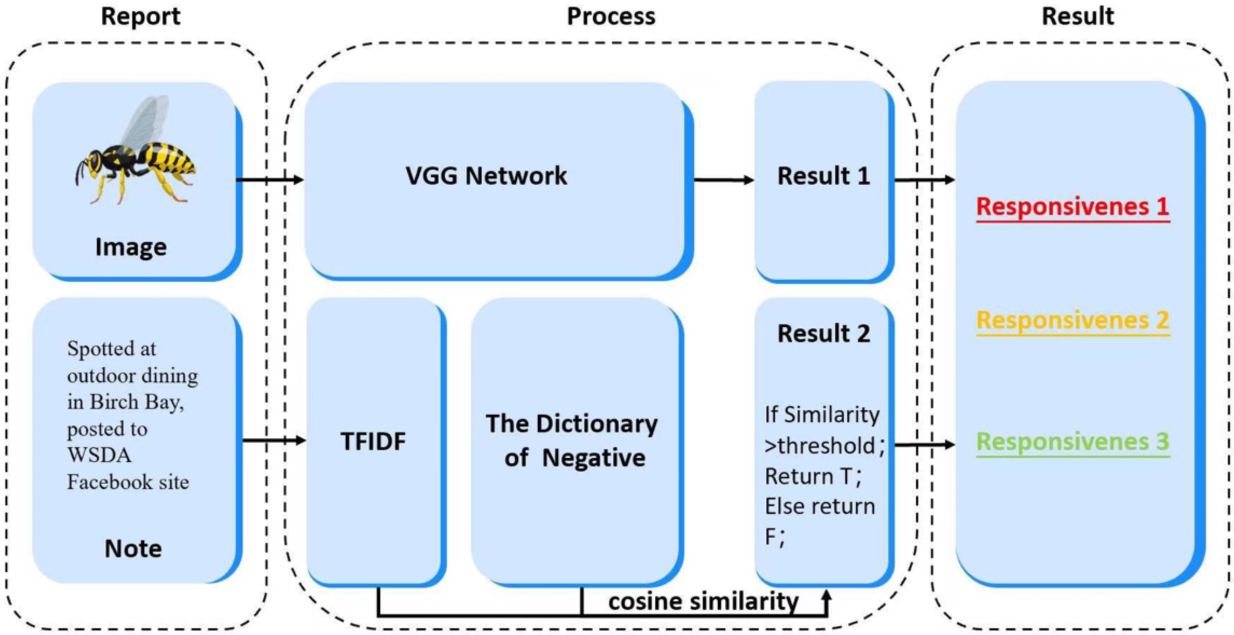

Report
Process
Result
Image
VGG Network
Result 1
Responsivenes 1
Spotted at outdoor dining in Birch Bay, posted to WSDA Facebook site
Note
TFIDF
The Dictionary of Negative
Result 2
If Similarity >threshold; Return T; Else return F;
Responsivenes 2
Responsivenes 3
cosine similarity

Figure 16: The structure diagram of the rapid response model

In the model, we use TF-IDF (term frequencyinverse document frequency) to extract word vectors. Tfidf is a commonly used weighting technique for information retrieval and data mining. TF is Term Frequency, and IDF is Inverse Document Frequency. It is used to evaluate the importance of a word to a document set or a document in a corpus. The importance of a word increases in proportion to the number of times it appears in the document, but it decreases in inverse proportion to the frequency of its appearance in the corpus.

We also use the cosine distance to calculate the similarity between strings. The cosine similarity calculation formula is as follows:

$$similarity = cos(\theta) = \frac{A \cdot B}{||A||\ ||B||} \tag{11}$$

A and B represents two word vectors in our model.

Specifically, when the wasp image detection model in problem two determines that the image contained in a certain report is a wasp, we will also judge the similarity between the note contained in this report and the negative dictionary string. Combining image recognition results and similarity, we designed a three-level response, as shown below:

- **First-level Response:** When the image and similarity both reach the threshold , we define it as the first-level response (that is, the report that should be processed first, most likely a bumblebee)

- **Second-level Response:** When either the image or the similarity reaches the threshold, we define it as the secondary response (that is, the report corresponding to the priority processing, which may be a bumblebee)
- **Third-level response:** When the image and similarity does not reach the threshold, we define it as a three-level response (that is, the report that does not need to be processed immediately, very unlikely a bumblebee)

The design of model threshold value will be given in detail in the problem solving.

3.4.3 Model solving

In the process of solving the fast response model, the key value is the text similarity threshold.We compare the similarity between positive text and negative dictionary strings . We also compare the similarity between negative text and negative dictionary strings. Consequently, we find that the former is much less similar than the latter.

Table 4: Similarity statistics table

GlobalID	Lat status	Similarty
5AC8034E-5B46-4294-85F0-5B13117EBEFE	Positive ID	0.371
2138197A-F5CF-4308-93E2-62EA6F84D098	Positive ID	0.155
DEF5D82B-E326-41A5-9B6CD46DCD86950C	Positive ID	0.225
C4F44511-EA53-4FCF-9422-E1C57703720D	Negative ID	0.894
89C867F1-D5ED-48C8-9586-B705F5DA9838	Negative ID	0.768
81670D96-4143-47B1-A9C8-83977892D53F	Negative ID	0.882
D30895B7-3994-45A3-BD51-E5BA881833FD	Negative ID	0.813

Through the table 4, we can clearly see the effectiveness of the similarity comparison. Taking into account the number of Positive IDs given in the question, we set the threshold to the mean value of the similarity of each Positive ID containing notes. We calculate the mean value of similarity of each Positive ID: 0.287. So the text similarity threshold is solved as 0.287.

After obtaining the threshold of the model, we filter out the unprocessed and unverified data and send it to the model to determine its response degree. Some of the first and second response degree reports are shown below.

Table 5: Responsibility judgment result table

GloballID	image	Simlarty	Responsiveness
AE7F3E03-F339-4081-8D83-CFCBC8B9DAA9	T	0.268	1
3E50801D-9DBB-43DE-8D32-31CFA88C74D9	T	0.253	1
BBBA5BA0-CAFB-43D3-8F1D-FB2D9CF777E0	T	0.847	2
41EB9554-6676-4CD1-BE5B-5133CB3EE4B0	T	0.687	2
016DDF49-2218-4073-A5F4-32396D0D4B70	T	0.777	2

The pictures corresponding to the first and second level reports of unprocessed and unverified data are as follows:

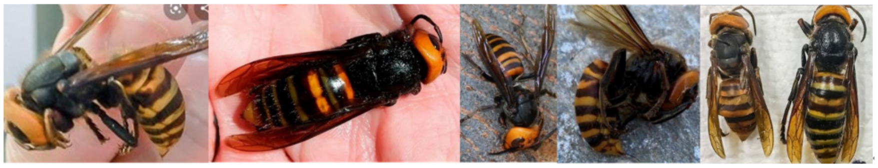


Figure 17: The picture corresponding to the report of the first and second response levels

3.4.4 Conclusion

From the model results , the model can be used as the credibility of whether the bee is an Asian giant hornet through the similarity and response degree .

3.5 Problem 4

3.5.1 Data preprocessing

Through consulting the literature and analyzing the information provided by this subject, we know that the appearance of wasps has obvious seasonal changes.Data show that the high frequency period of Asian bumblebee is from May to August.Other similar species of bees appeared around this time.This will lead to an increase in the number of reports.To better determine the frequency of model updates, we will first count the frequency of reports and the relationship between the months.

The data given contains a lot of invalid date information, while the month and season factors are important information that needs attention. For more detailed analysis, we clean the invalid date data and combine the number of reports appearing in the same month in different years as well. The relationship between the frequency of report and the month is shown in the following table:

Table 6: Frequency statistics in different months

Month	Reports	Negative	Positive	Unprocessed	Unverified
1	6	2	0	0	4
2	5	3	0	0	2
3	23	5	0	0	18
4	194	44	0	0	150
5	581	129	2	0	450
6	342	187	1	0	154
7	1028	542	0	0	486
8	1396	715	1	6	674
9	660	353	6	2	299
10	183	87	2	7	87
11	5	0	1	1	3
12	3	0	1	0	2

3.5.2 Problem Analysis and Modeling

In this problem, we need to figure out the update frequency of the model. First, we consult relevant literature and learn that the deep learning model used in Question 2 requires a large amount of computing resources and data. Nevertheless, the real world situation is more complicated and changeable. So outdated models often fail to achieve excellent results. Based on the above reasons, we have the most significant principle of our model updating: to update the model as quickly as possible under the premise of ensuring the effective data volume.

From the relationship between the frequency of the report in the data preprocessing section and the month, we can see that the report is concentrated in April to October, which also corresponds to the habitable month of the wasp. Simultaneously, the relationship between frequency and month fits the Gaussian normal distribution function. The figure is following below

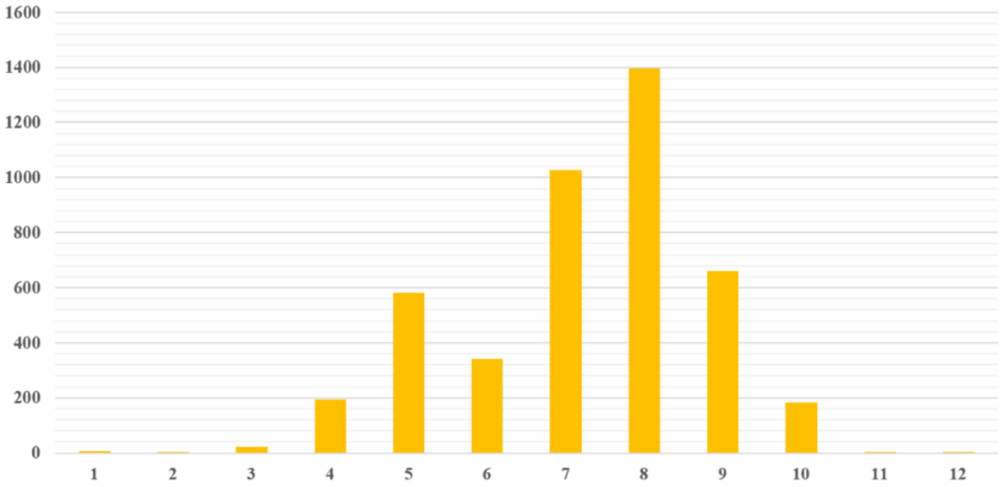


Figure 18: Frequency and monthly statistics

From the figure we can see that the report often peaks between August and September. A large number of reports correspond to a large amount of data, which can satisfy the training of our model.

3.5.3 Conclusion

We conclude that our model should be updated between August and September each year, and the update frequency should be once a year. It can not only ensure the timeliness of the model, but can also satisfy the data amount required by model training.

3.6 Problem 5

According to the conclusions of questions 1 to 4, the criteria for determining the pests are determined as follows: after the sighting report is updated, the new data is determined by the model to no longer have the primary and secondary responses, indicating that Washington State has got rid of the pest haze.

4 Strengths and Weaknesses

4.1 Strengths:

- In the modeling of the second problem, we use the method of data augmentation to compensate for the imbalance of positive and negative samples. On this basis, we propose a wasp image recognition model derived on a convolutional neural network. Compared with traditional MLP, convolutional neural network weight sharing in the convolutional layer in CNN reduces the number of trainable parameters in the network and the complexity of the network model. It also has a less overfitting phenomenon. Therefore, it obtains a better generalization ability.

  At the same time, the pooling operation used in CNN significantly reduces the number of neurons in the model and is more robust to the translation of the input space.

  In addition, the introduction of migration learning allows our model to achieve better classification results on small data sets.

- We propose a set of artificial intelligence-based wasp governance systems in our memo of advice to the Washington government.

- In the third question, we comprehensively consider the multi-modal information of text and image. Consequently, we advocate a rapid response model based on multi-modal information. This model comprehensively considers the text and image information in the report. Compared with using only a single modal information, our model builds three response levels based on the results of multi-modal information feedback in order to respond to the report more objectively.

4.2 Weaknesses:

- The wasp spatial position prediction model in the first model uses less data for training. In contrast, if there is enough data, the LSTM-based wasp spatial position prediction model will achieve better results.

- Although the wasp image recognition model based on the convolutional neural network in model two has achieved good results in the field of identifying wasps, the convolutional neural network itself requires higher computer resources and longer training time than traditional methods.

5 Memorandum

To: the Washington State Department of Agriculture
From: Team #2113265
Date: 9 February 2021
Subject: Perfecting the Government's Elimination of Vespa Mandarine

Dear the Washington State Department of Agriculture:

It is our honour to help you formulate policies on Vespa Mandarine to actively respond to pest control. In order to facilitate you to carry out more accurate investigations and protect local biodiversity and social ecological stability, we are very willing to share our modeling results and attach some related policy recommendations based on the models.

We build a three-part model . The first part of the model predict the spread of Vespa mandarine in Washington State over time . The latter two parts of the model establish a three-level response mechanism for public reports by processing image and text data . This mechanism can assess unprocessed sighting reports on three response levels . The results of the response level can be used as the basis of you to formulate the report investigation priority .In the process of building the model, we analyze the seasonal data to determine the active season of the bee colony , which is used to determine the degree of control work carried out at different times . In order to eliminate pests and reduce the serious impact of species invasion on the ecology and economics of the western United States, we throw out the following six suggestions:

1 We use the model to predict the spread of Vespa Mandarine over time and assess the risk level of being affected by pests under natural environmental conditions. It is recommended that the forecast ranges and statistics provided here be used in conjunction with the pest response guidelines issued by the United States Department of Agriculture so that state governments at all levels can develop strong control strategies.

2 We use the model to judge the reports that have not yet been determined and not yet processed. In 2344 cases, it is determined that there are five reports with higher response levels. Two of them were first grade and three were second grade. We believe that these five reports are likely to be potential positive IDs. It is recommended that the five ones should be given priority. Careful monitoring and investigation should be carried out within two kilometers of the reported location, and the location of the nests should be tracked by hormones and other biological means for the pests that are found. Do not forget to destroy their nests.

3 Optimize existing websites and platform hotlines for collecting public reports and improve the standardization and accuracy of collected data, so as to regulate and improve the accuracy of these public reports.

4 In May of each year, use the model to evaluate the new sighting report, as well as use the grade result determined by the model as the basis for the investigation. May is the most active period in the life cycle of the swarm. You can find the nest of Vespa Mandarine more efficiently and eliminate it before expanding the scale of development. After a certain number of new Positive

IDs are obtained through the above work, they are imported into the model as new basic data for analysis. This approach can get a more accurate prediction map.

5 We evaluate the pest risk level in Washington State through the model. We believe that if you update the data, the new sighting report is judged by the model and no longer has the primary and secondary responses, indicating that Washington State has got rid of the pest haze.

We believe that the model and data we provide will contribute to the response to this species and provide help for the rapid assessment of the priority of numerous public reports in the future. We hope that our model can perfect the government's elimination of pests reducing people's panic as well as returning the natural ecology of Washington State to its original state. So that's the summary of our study. We sincerely hope that it will provide you with useful information. Thank you very much!

References

[1] Aversano L , Bernardi M L , Cimitile M , et al. Tomato diseases Classification Based on VGG and Transfer Learning[C]// 2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor). IEEE, 2020. Publishing Company , 1984-1986.

[2] Yang F W , Lin H J , Yen S H , et al. A Study on the Convolutional Neural Algorithm of Image Style Transfer[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2018.

[3] Jing L P , Huang H K , Shi H B . Improved feature selection approach TFIDF in text mining[C]// International Conference on Machine Learning & Cybernetics. IEEE, 2002.

[4] Van Lenteren J C , Woets J . Biological and integrated pest control in greenhouses.[J]. Annual Review of Entomology, 1988, 33(1):239-269.

[5] Gers, Felix A , Schmidhuber, et al. Learning to Forget: Continual Prediction with LSTM.[J]. Neural Computation, 2000.

[6] Zheng Z , Ahn S , Chen D , et al. Freeway traffic oscillations: Microscopic analysis of formations and propagations using Wavelet Transform[J]. Transportation Research Part B, 2011, 45(9):1378-1388.

[7] Risch S J , Andow D , Altieri M A . Agroecosystem Diversity and Pest Control: Data, Tentative Conclusions, and New Research Directions[J]. Environmental Entomology, 1983, 12(3):625-629.

[8] Jianping Ju,Hong Zheng,Xiaohang Xu,Zhongyuan Guo,Zhaohui Zheng,Mingyu Lin. Classification of jujube defects in small data sets based on transfer learning[J]. Neural Computing and Applications,2021

[9] D. Lakshmi,K. Palani Thanaraj,M. Arunmozhi. Convolutional neural network in the detection of lung carcinoma using transfer learning approach[J]. International Journal of Imaging Systems and Technology,2020,30(2).

[10] Venkatesan Rajinikanth,Alex Noel Joseph Raj,Krishnan Palani Thanaraj,Ganesh R. Naik. A Customized VGG19 Network with Concatenation of Deep and Handcrafted Features for Brain Tumor Detection[J]. Applied Sciences,2020,10(10).

[11] Mindy Ward. Dectes stem borer trapping back on in Missouri[J]. Southwest Farm Press,2021.

Appendices

Here are simulation programmes we used in our model as follow.

Appendix A

Python code for image training classification

```
##Image classification model training code
from keras import optimizers
from keras import applications
from keras.models import Sequential, Model
from keras.callbacks import ModelCheckpoint
from keras.layers import Dropout, Flatten, Dense, GlobalAveragePooling2D
from keras.preprocessing.image import ImageDataGenerator
img_height, img_width = 256, 256
batch_size = 32
epochs = 20
train_data_dir = ’./train’
validation_data_dir = ’./val’
OUT_CATEGORIES = 1
base_model = applications.VGG16(weights="imagenet", include_top=False,
                               input_shape=(img_width, img_height, 3))
print(base_model.summary())
train_datagen = ImageDataGenerator(rescale=1. / 255, horizontal_flip=True)
test_datagen = ImageDataGenerator(rescale=1. / 255)
train_generator = train_datagen.flow_from_directory(train_data_dir,
   target_size=(img_height, img_width),
                                                     batch_size=batch_size,
                                                         class_mode=’binary’)
validation_generator = test_datagen.flow_from_directory(validation_data_dir,
   target_size=(img_height, img_width),
                                                       batch_size=batch_size,
                                                           class_mode=’binary’,
                                                           shuffle=False)
for layer in base_model.layers[:15]: layer.trainable = False
top_model = Sequential()
top_model.add(Flatten(input_shape=base_model.output_shape[1:]))
top_model.add(Dense(256, activation=’relu’))
top_model.add(Dropout(0.5))
top_model.add(Dense(OUT_CATEGORIES, activation=’sigmoid’))
model = Model(inputs=base_model.input, outputs=top_model(base_model.output))
model.compile(loss=’binary_crossentropy’, optimizer=optimizers.SGD(lr=0.0001,
   momentum=0.9),
              metrics=[’accuracy’])
checkpointer = ModelCheckpoint(filepath=’dogcatmodel.h5’, verbose=1,
   save_best_only=True)
history=model.fit_generator(train_generator, steps_per_epoch=nb_train_samples
   // batch_size, epochs=epochs,
                    validation_data=validation_generator, validation_steps=
                        nb_validation_samples // batch_size,
                    verbose=1, workers=12, callbacks=[checkpointer])  # 12
import pickle
with open(’log.txt’, ’wb’) as file_txt:
    pickle.dump(history.history, file_txt)
```

Appendix B

Python code for LSTM training

```
def NormalizeMult(data, set_range):
    normalize = np.arange(2*data.shape[1], dtype=’float64’)
    normalize = normalize.reshape(data.shape[1], 2)
    for i in range(0, data.shape[1]):
        if set_range == True:
            list = data[:, i]
            listlow, listhigh = np.percentile(list, [0, 100])
        else:
            if i == 0:
                listlow = -90
                listhigh = 90
            else:
                listlow = -180
                listhigh = 180
        normalize[i, 0] = listlow
        normalize[i, 1] = listhigh

        delta = listhigh - listlow
        if delta != 0:
            for j in range(0, data.shape[0]):
                data[j, i] = (data[j, i] - listlow)/delta
    return data, normalize
def trainModel(train_X, train_Y):
    ’’’
    trainXtrainY: Data needed to train the LSTM model
    ’’’
    model = Sequential()
    model.add(LSTM(32,return_sequences=False))
    model.add(Dropout(15))
    model.add(Dense(train_Y.shape[1]))
    model.add(Activation("relu"))
    model.compile(loss=’mse’, optimizer=’adam’, metrics=[’acc’])
    model.fit(train_X, train_Y, epochs=15, batch_size=4, verbose=1)
    model.summary()
    return model
```