A Preliminary Intergenerational Photo Conversation Support System based on Fine-tuning VGG16 Model

Lei JIANG[1], Panote Siriaraya[2], Noriaki Kuwahara[4]
Graduate School of Science and Technology
Kyoto Institute of Technology
Kyoto, Japan

Dongeun Choi[3]
Faculty of Informatics
The University of Fukuchiyama
Kyoto, Japan

***Abstract*—China has the largest number of elderly people in the world, young volunteers have become the main force in caring for the elderly. It is urgent to establish a photo conversation support system to build a bridge of communication between young volunteers and the elderly. Previous research generally used perceptual analysis or machine learning methods to find photos suitable for intergenerational conversation, this paper uses deep learning models to further learn the potential features of two datasets suitable for and not suitable for intergenerational conversations. However, the original datasets are too small, it was first proposed to use TF-IDF in conversation recording and data augmentation technology in images to expand the datasets. Then on the basis of the VGG16 model combined with transfer learning and fine-tuning technologies, five models were designed. The accuracy of the best model on the validation set and test set reached 96% and 94.5%. In particular, the recall rate of the not suitable dataset reached 100%, all not suitable photos were identified. At the same time, the recall rate of other datasets reached 71% for not suitable photos. It shows that the system is also applicable to other datasets and can effectively eliminate photos that are not suitable for intergenerational conversations.**

Keywords—Intergenerational photo conversation support system; TF-IDF; VGG16; transfer learning; fine-tune

I. Introduction

A. Changes in Care Models and Cognitive Impairment

China has the largest aging population in the world. We urgently need to address the problem of improving quality of life for these senior citizens. , Currently, there are three modes of family care in China: institutional care and community home care [1]. As the proportion of 4-2-1 structured households (four seniors, a couple and one child) increases year by year, couples have little energy to allocate to the elderly, and it is difficult to meet the emotional needs of the elderly on a purely material basis [1]. Although institutional old-age care promotes the “integration of medical and nursing care” [2], and the companionship of other elderly people can reduce loneliness, China's experience in changing from an adult society to an aging society has been relatively fast, and the number of places is insufficient. On average, there are only 21.5 places per 1,000 senior citizens [3]. Community home care is provided by the community based around the family being the core provider of related care services for the elderly living at home. It is an effective long-term care method that is popular among the elderly. However, there are also some problems. There is a shortage of professional personnel for care giving, and the majority of the workforce is made up of volunteers. As of 2012, 33.92 million people had registered as youth volunteers[4]. However they only account for 2.54% of the total population and few people regularly participate in voluntary activities for a significant amount of time, mostly due to lack of communication and emotional exchanges[4]. If the volunteers lack experience in caring for and communicating with the elderly, it can become difficult for the two generations to empathize and find common ground. At the same time, the increase of non-traditional family structures such as dink families (double income and no kids), lost families, and empty nest families means that for 47.53% of the elderly in China in 2016 are in these non-traditional family structures[5]. 60% of the elderly in non-traditional family structures have psychological problems related to cognitive impairment[6]. Strengthening daily communication to ensure good social relations and minimizing psychological distress may help delay or prevent cognitive impairment[7].Therefore, it is extremely important to find a way to help young people communicate with the elderly and help them to establish emotional connections.

B. Research Purpose

This paper aims to establish a preliminary intergenerational conversation support system which is able to help the elderly and young volunteers to quickly and effectively screen their favorite photos to inspire conversation.

C. Research Materials

In the early stage, our research team has done two generations of conversation experiments supported by photos, and grouped all the photos into groups that are suitable for the conversation of the elderly and not suitable for the elderly. The research in this paper is based on the results in [8]. Through the deep learning model to learn the characteristics of each group of photos. So that it can effectively eliminate photos that are not suitable for two-generation conversations and obtain suitable photo.

The first group of photos are the third cluster of [8] which are suitable for the elderly to talk about. It is classified as the good category that inspired many conversation topics and made conversation flow freely between the two generations. These contents are: “cakes”, “lotus flowers”, “shared bicycles”, “red dates and white fungus soup”, “pandas”, “foot therapy”, “movie theaters”, “Water Cube (Beijing National Aquatics

Center)”, and “outings”. The second group of photos are the first cluster of the paper which are not suitable for the two generations communication. It is classified as the bad category, as the content of the conversation is limited and both parties are stressed. These contents are: “Shengjianbao”, ”Zhang Yimou and Gong Li”, “Qinhuai River”, “mother teaches children to fold clothes”, “Jackie Chan”, “toothpaste”, “rape flowers”, and “gourds”.

Although the experimental photos are divided into two groups, it is difficult to distinguish the category of each group of photos by traditional statistical methods. For example, “raw fried buns”, “cakes”, “red dates and white fungus soup” all belong to the food category, but the elderly and young people have different attitudes towards them. “Rapeseed” and “lotus” both belong to plants, but the attitudes towards them are also different. Only after conducting enough dialogue experiments can we understand the relationship between the groups and the photos within the groups. This process is very time-consuming and labor intensive. In order to overcome this difficulty, we chose to use deep learning models to help us learn the potential relationships between and within groups, provide us with guidance to choose photos for the next set of conversations, and to initially establish a photo conversation support system. The main contributions of this paper present the following:

- In order to solve the problem of insufficient data set for the deep learning models , not only the commonly image processing techniques used in photos such as rotation and stretching, but also the TF-IDF technology used to expand the content of the photos to achieve the expansion of the data set purpose.
- On the basis of the VGG16 model, five models combined with transfer learning and fine-tuning technologies were designed. Compared with the CNN model trained from scratch, it greatly improves the accuracy and reduces the amount of calculation and training time.
- Data sets from other regions are used to explore the applicability of the model of this system.

Rest of the paper is organized as follows. Section II introduces an overview of the existing research. Section III presents data acquisition, date augmentation, model architectures, model evaluation and the process of establishing the intergenerational photo conversation support system. Results obtained the model which is able effectively eliminate photos that are not suitable for two-generation conversations in Section IV. Section V presents the discussion of this paper. The conclusion and perspectives of this paper are provided in Section VI.

II. Related Work Reviewer A Q3

Astell et al. [9-12] developed CIRCA, a multimedia system displayed on a touch screen for dementia patients. The project started in 2001 and now includes photos, videos, music, graphics and text. CIRCA is designed to be used by dementia patients and caregivers, and it provides cognitive support for people with dementia covering communication, entertainment and creativity. Alm et al. [13] carried out a long-term trial and evaluation of CIRCA, expanded the multimedia content and introduced a randomization function, and found that this was able to provide enough interesting content to stimulate dementia patients to recall memories that were previously unheard. CIRCA was developed in Dundee to explore the possibility of using the system in other regions, and Purves et al. [14] selected relevant content for seniors for a British Columbia version of CIRCA. The pilot tested CIRCA-BC on 3 participants with dementia and a conversation partner. By analyzing their interactions, they found that CIRCA's content can be adjusted for use in different regions based on similarities and differences in the participants' shared social history. Fels et al. [15] proposed using everyday photos to inspire dementia patients to share stories about their lives and life experiences in order to establish emotional connections. He found that regular daily conversations allow participants and listeners to have pleasant shared interactions. Miyuki Iwamoto et al. [16] researched a dialogue support system hoping to reduce the stress of young volunteers in the system when talking to the elderly through the shared content of photos and videos. The results of this were that the video content caused subjects to feel less stressed, but that the photos allowed conversation between older people and young volunteers to last for longer periods of time. Relatively little research has been carried out in China. Zhou et al. [17] took 40 photos and conducted 160 conversation experiments. He discovered that photos of “food”, “event”, “school” and “commodity” were suitable for conversation between two generations of Chinese. Zhou in another paper[8] used the PCA method and HCA to analyze and classify 40 photos, hoping to find the best photo cluster suitable for establishing an intergenerational photo dialogue support system. Although this research did not achieve a perfect result, it provided guidance for photo classification projects in the future.

From previous related researches, three main points can be concluded. The one is that photos make the talker more focused and more relaxed and can stimulate participants to share stories of personal life and life experience. so we used photos as the material for the intergenerational conversation is feasible. The other is that conversation system can be applied to different regions, if the materials are carefully selected. Therefore, our research in three different cities were conducted to explore the possibility of applying the intergenerational photo conversation support system in different regions. The last, previous researches generally used perceptual analysis or machine learning methods to find out the photos suitable for conversation and the deep learning models have never been used.

III. Research Methodology

A. Data Acquisition

Each group contains very few photos and as such, we cannot use deep learning models. We can instead expand the semantics of photos through natural language technology. Specifically, each photo contains four groups of dialogues, and we perform word segmentation and keyword extraction on the dialogues. We are able to use the keywords as candidate words to design a questionnaire, in order to first determine the topics of the candidate words that the two generations are and are not interested in, and then to select photos according to each of these topics. Finally, two types of photos are counted, and the photos in each category are randomly divided into a training set and a verification set at a 2: 1 ratio.

1) Keyword extraction: In order to dig deep into the topics that old and young people like and dislike, we analyzed their conversation content for each photo through TF-IDF and extracted the keywords. TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and data mining [18]. It uses numerical statistics to reflect the importance of a word relative to a document in the corpus[19]. The first step is to calculate the TF:

$$\mathrm{TF} = \frac{\mathrm{count(t)}}{\mathrm{count}(\mathrm{d_i})} \tag{1}$$

In the equation(1), count (t) represents the number of words contained in the document; count (d_i) represents the total number of words in the document. As can be seen above, if a word appears frequently in a document the value of TF is high.

The second step is to calculate the IDF:

$$\mathrm{IDF} = \frac{\mathrm{num(corpus)}}{\mathrm{num(t)+1}} \tag{2}$$

In the equation(2), num (corpus) represents the total number of documents in the corpus; num (t) represents the number of documents in the corpus containing the particular words. As can be seen above, if a word rarely appears in other documents in the corpus, the value of IDF is low.

The third step is to calculate the TF-IDF:

$$\mathrm{TF—IDF} = \mathrm{TF(t)} \times \mathrm{IDF(t)} \tag{3}$$

In the equation(3), the higher the TF-IDF of a word, the more it is considered to be important and representative for the documents. The TF-IDF algorithm can be implemented in any programming language of your choice. Because the Chinese document must be preprocessed using word segmentation and stop word filtering, before the TF-IDF calculation, this article uses a module named jieba with python for convenience. We calculated the TF-IDF value of the keywords of each photo in descending order, and selected the top five.

2) Data selection by questionnaire: Questionnaires were designed to confirm effective keywords and further explore the opinions of older people towards the photos. Table I shows the questionnaire participants. A total of two questionnaires were designed. Questionnaire one was used to let elderly people select keywords with positive impressions from fifty keywords that were extracted from the good category and then select positive pictures related to each keyword selected. Finally, the younger participants choose their positive pictures from the results of the elderly group. Questionnaire one includes two parts. The first ,for the elderly:

1. Do you have a positive impression of this keyword?

– Yes, I do. Please see question 2.

– No, I don’t. Please see the next keyword.

– I don't know. Please see the next keyword.

2. Do you like this photo related to the keyword?

– Yes, I do. (Keep this picture)

– No, I don't. (Delete this picture)

– I don't know. (Delete this picture)

The another for the young: Do you have a positive impression of this photo (Photos selected by the elderly group)?

– Yes, I do. (Keep this picture)

– No, I don't. (Delete this picture)

– I don't know. (Keep this picture)

Questionnaire two is only for seniors. The elderly person selects keywords with negative impressions from forty keywords extracted from the bad category and then selects the negative pictures related to each keyword selected. The content of the questionnaire two for the elderly:

1. Do you have a negative impression of this keyword?

– Yes, I do. Please see question two.

– No, I don't .Please see the next keyword.

– I don't know. Please see the next keyword.

2. Do you dislike this photo related to the keyword?

– Yes. I have. (Keep this picture)

– No, I don't.(Delete this picture)

– No, I don't know(Keep this picture)

In this way, there are no photos in the photo collection that the elderly do not like, and all of the photos are liked by both the elderly and young people.

B. Data Augmentation

All photos collected were divided via the questionnaires into a training set and a validation set in a 2: 1 ratio of the two categories, good and bad. We then augmented the training set using the keras preprocessing module ImageDataGenerator in python. There are many options available in ImageDataGenerator, but we chose just four of them and set Rescale = 1./255, Shear_range = 0.2, Zoom_range =0.2, Horizontal_filp = True. By using this tool our training sets were amplified tenfold. Rescale is a value by which we can multiply the data before any other processing[20]. Our original images consist of RGB coefficients in the range 0-255, so we target values between to 0 and 1 to instead by scaling with a factor of 1./255. Shear_range is for randomly applying shearing transformations. Zoom_range is for randomly zooming into pictures. Horizontal_filp is for randomly flipping half of the images horizontally relevant when there are no assumptions of horizontal asymmetry.

TABLE I. THE PARTICIPANTS OF QUESTIONNAIRE

No.	Questionnaire Location	Participant	Dementia
1	Shanghai Qibao Vanke City Garden Community	2 elderly females and 3 elderly males	No
2	Shanghai Qibao Vanke City Garden Community	2 young females and 1 young male	No

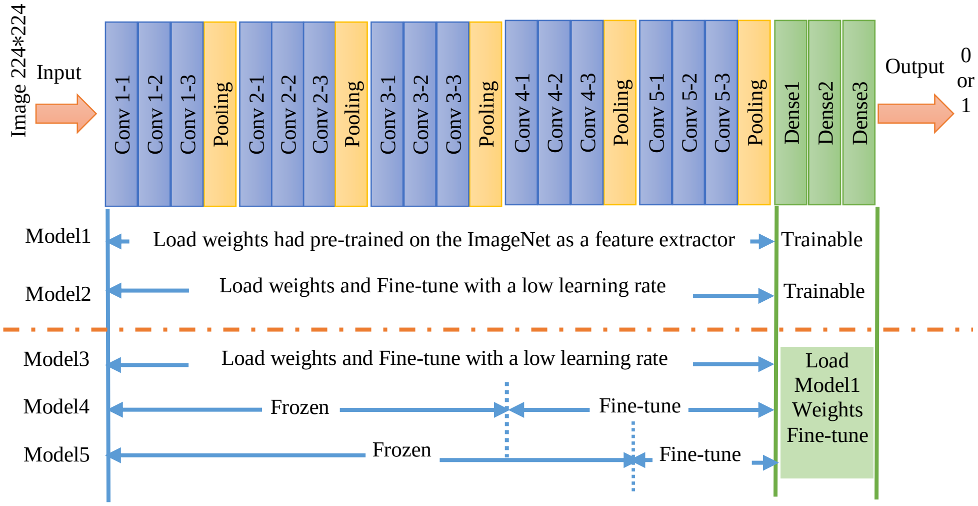


Fig. 1. VGG16 Model Architecture and Designed Models Architecture.

C. Model Design and Architecture

Although photos have been augmented on the content and number of photos, the data set is still too small for a deep learning model. For small amounts of data, data augmentation, transfer learning, fine-tuning, or a combination of several methods is generally used as in [21-24]. It used rotation, skewing and elastic distortion augmentation methods for images and then used a pre-trained CNN model as feature extractor and SVM as a category classifier. This technique has been applied in many fields and has shown better accuracy than traditional convolutional networks [25-27]. In this paper after we augmented the training set, we combined transfer learning and fine-tuning and produced five models (Fig. 1), all of which were implemented through the keras framework of python3.

The VGG16 architecture was selected to transfer learning and its weights were fine-tuned with a low learning rate. VGG16 with 13 convolutional layers and 3 fully connected layers. The first time, after two convolutions of 64 convolution kernels, one pooling is used. The second time, after two 128 convolution kernel convolutions, pooling is used. Repeat the convolution with three 512 convolution kernels twice, and then pooling. Finally, after three full connections. Because it uses a 3×3filter with a stride of 1 to construct the convolutional layer, the padding parameter is the parameter in the same convolution. Then use a 2×2 filter with a stride of 2 to construct the pooling layer, so that the VGG16 network does not have so many hyperparameters. Therefore, one of the advantages of the VGG network is that it does simplify the structure of the neural network, but it is very deep and can learn more deep features.

At the same time VGG16 model had pre-trained on the ImageNet dataset containing 1000 classes which already had learned features that are relevant to our classification issues. The details are as follows (Fig. 1), Model 1: Load the weights of pre-trained VGG16 as a feature extractor and then train full connection layers as their own category classifier; Model 2: loading the weights of pre-trained VGG16 and all layers fine-tuned with a low learning rate; From Model 3 to Model 5: Load the pre-trained VGG16 weights and Model 1 fully-connected layer weights, then freeze different convolutional layers (fixed parameters) and fine-tune the unfrozen layers.

Dense1 layer is a GAP (Global Average Pooling) layer with dropout. GAP was first proposed in and is considered a new technology that can replace the fully connected layer, especially in transfer learning[28, 29]. Assuming that the final output of the convolution layer is a three-dimensional feature map of h × w × d, the specific size is 6 × 6 × 3, after GAP conversion, it becomes an output value of size 1 × 1 × 3, that is, each of the layers h × w will be averaged to a value. There are two advantages, one is that GAP is more simple and natural to convert between the feature map and the final classification. The second is that unlike the FC layer that requires a lot of training and tuning parameters, reducing the spatial parameters will make the model more robust and better resist overfitting. Using our data set, comparing the GAP, GMP and FC layers, we found that the GAP technology was relatively stable in accuracy and loss rate, and the effect of anti-overfitting was obvious. Dropout technology was first proposed in [30]. Later, it was applied to CNN network models for image classification [31]. The effect of anti-overfitting is obvious. The principle is to discard some neurons with probability P during transmission, other neurons are retained with probability q = 1-

p, and the output of the discarded neurons is set to zero [32]. In this article, P was chosen to be 0.2(0.1,0.2,0.5 were tested). Dense2 layer is a full connection layer with 100 neurons and ReLu as activation function.Dense3 layer using Softmax function with two neurons achieving two classification effect.

All images were resized into 224×224 then input into the per-trained model (Model 1 to Model 5). The hyper parameters of the model were selected: learning rate to 0.0001(0.01, 0.001 were tested), batch size to 64, number of epochs to 20, loss was categorical_crossentropy, Adam optimizer (Rmsprop, SGD, Adam were tested).In order to prove the effectiveness of the designed models in this paper, a CNN model with three convolution layers with a ReLU activation and followed by max-pooling layers to train it from scratch as a benchmark for comparison. All images were resized into 224×224 then input into the CNN. The hyper parameters were: learning rate to 0.001, batch size to 32 number of epochs to 100, Adam optimizer, loss was categorical_crossentropy.

D. Model Evaluation

The confusion matrix is an indicator to judge the result of the model, and is generally used to judge the quality of the classifier [33]. The confusion matrix can reflect the effectiveness and accuracy of the model in more detail than the evaluation index. We can clearly see the identification of each type of sample. Here the good category was defined as the label 0 and the bad category as the label 1. For the label 0 : the number of good category photos correctly identified as label 0 by the model is a, the number of label incorrectly identified as label 1 is b; For the label 1: the number of correctly identified as label 1 by the model is c, and the number of incorrectly identified as label 0 by the model is d. Therefore, the form of confusion matrix as follow:

$$\text{Confusion Matrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad (4)$$

The following evaluation indicators are all based on the confusion matrix, specific calculation process is as follows:

- In the equation(5) and (6), precision (Accuracy rate) means the proportion of the good(bad) category photos that are actually classified as label 0(1). In general, the higher the accuracy, the better the model.

$$\text{Precision}(0) = \frac{a}{a+c} \quad (5)$$

$$\text{Precision}(1) = \frac{d}{b+d} \quad (6)$$

- In the equation(7) and (8), recall represents the ratio of the number of good(bad) photos that the model identified as label 0(1) to the total number of good(bad) category dataset. In general, the higher the Recall, the more good(bad) photos are predicted by the model.

$$\text{Recall}(0) = \frac{a}{a+b} \quad (7)$$

$$\text{Recall}(1) = \frac{d}{c+d} \quad (8)$$

- In the equation (9) and (10), F1_Score which is defined as the harmonic average of correctness and recall. The value of F1-Score is from 0 to 1, with 1 being the best and 0 being the worst.

$$F1_{Score(0)} = \frac{2Precision(0)*Recall(0)}{Precision\ (0)+Recall(0)} \quad (9)$$

$$F1_{Score(1)} = \frac{2Precision\ (1)*Recall(1)}{Precision(1)+Recall(1)} \quad (10)$$

- The abscissa of the ROC curve is the false positive rate (FPR is equal to recall(1)) and the ordinate is the true positive rate (TPR is equal recall(0)). AUC (Area under Curve) is defined as the area under the ROC curve, which clearly indicates which classifier has the better effect [31]. The classifier with larger AUC is better: 0.5 <AUC <1, which is better than random guessing. If the classifier properly sets the threshold, it can have predictive value; AUC = 0.5, like the follower guess, the model has no predictive value; AUC <0.5, which is less accurate than random guessing [34].

E. Applicability of the Model to Other Data Sets

Forty photos in the paper[17] were analyzed by the method described in [8]. 14 photos grouped as bad category which were “Blackboard”, “Classroom”, “Textbook”, “Dorm room”, “School”, “School bag”, “Diploma”, “Coal”, “Rice balls”, “Dumplings”, “Spring Festival Couplets”, “Zhongshan Mausoleum”, and “Ping pong”.

There were 16 photos grouped as good category, “Popcorn”, “Sausage”, “Ravioli”, “Meatballs with soy sauce”, “Fried rice”, “Hot pot”, “Spring Festival Gala”, “New Year”, “Dragon Boat Race”, “Children's day”, “Black and white TV”, “Tea bottle”, “Vintage bicycle”, “Kerosene lamp”, “Vanishing cream”, and “Quilt”. Fig. 4 shows the distribution of photos. All of them were inputted into the model for prediction to evaluate how well the model for other data.

F. Establishment of the Photo Conversation Support System

Finally a preliminary intergenerational photo conversation support system was established as shown in Fig. 2.

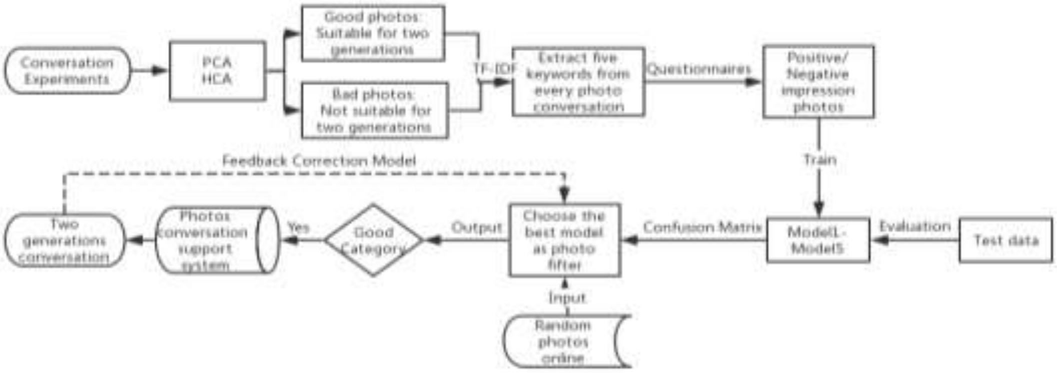
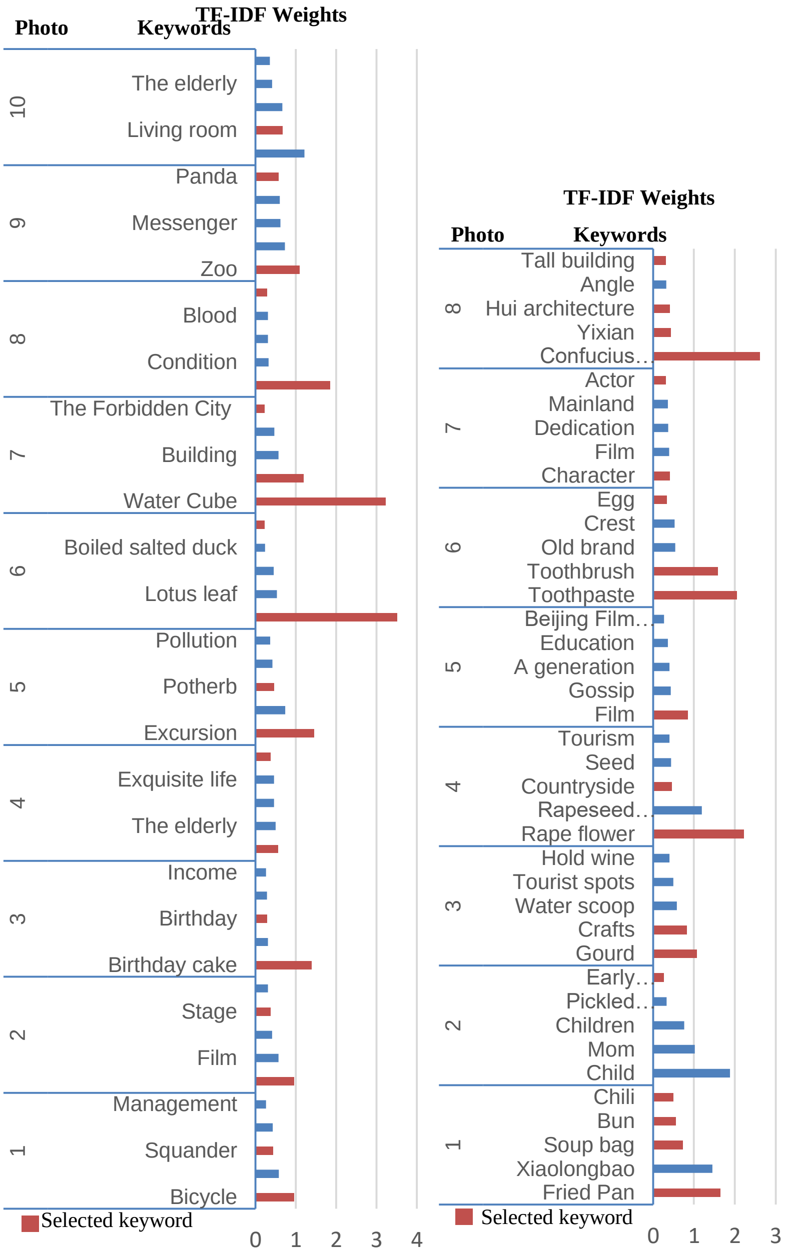

Conversation Experiments
PCA HCA
Good photos: Suitable for two generations
Bad photos: Not suitable for two generations
TF-IDF
Extract five keywords from every photo conversation
Questionnaires
Positive/ Negative impression photos
Train
Feedback Correction Model
Two generations conversation
Photos conversation support system
Yes
Good Category
Output
Choose the best model as photo filter
Confusion Matrix
Model1-Model5
Evaluation
Test data
Input
Random photos online

Fig. 2. Photo Conversation Support System.



(a) Positive Impression Photos

(b) Negative Impression Photos

Fig. 3. The Result of Keyword Extraction and Questionnaires.

IV. Results

A. Dataset Division

Records of conversations for each photo were extracted top five keywords by the TF-IDF technology. Then through questionnaires, each group selected twenty keywords by the elderly and the youngers. As shown in Fig. 3, every selected keyword had seventy-five photos selected by the young and the elderly.

Therefore, we have two photo categories: those with a positive impression (good category) and those with a negative impression (bad category), each with 1500 photos. Next, the photos in each category were randomly divided into a training set and a validation set at a 2: 1 ratio. The training set was 1000 photos and the validation set was 500 photos per category. The original 18 pictures were used as a test set, ten of which were in the good category and eight of which were in the bad category.

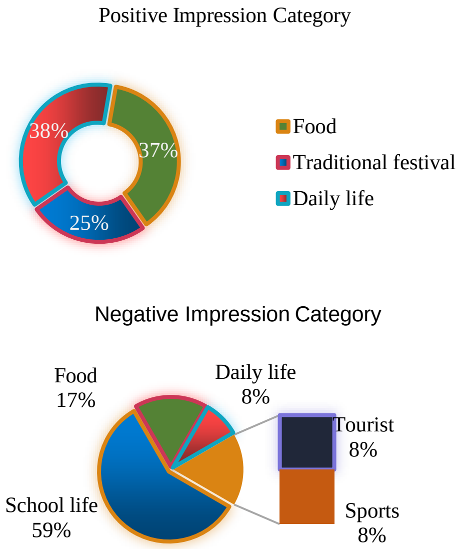


Fig. 4. The Distribution of Photo Categories.

B. Photo Classification Model

Table II shows that transfer learning and fine-tuning the network is more effective than training from scratch (CNN) for small data sets. The accuracy rate increased from 80% to over 94%. Compared to Model 1 using VGG16 as a feature extractor and training a fully connected layer as a classifier, Models 2 to 5 fine-tuned the VGG16 network to obtain better results. The best result was the accuracy rate increasing from 94.8% to 99.8%. And Model 3 and Model 4 have the highest accuracy on the training set and validation set. However, it is not important whether it identifies good category photos as bad category photos, as it will not affect our selection of photos for dialogue. The important thing is that we must not identify bad category photos as good category photos, so a higher rate of bad classification recognition was needed. To achieve this, the confusion matrix of the validation set and the test set used to further evaluate the model (Table III). From the results in Table III, it shows that Model 3 and 4 had the highest recognition rates (the recall value of the bad category) for the test set, but in the Validation set Model 4 was better than Model 3 (the classifier with larger AUC is better). Therefore Model 4 was finally selected.

C. Optimization of Model Parameters

In order to further improve the model, we increased the number of neurons in the Dense2 layer (Table IV). We found that in Model 4 every bad category photo was recognized when we applied a Dense2 layer with 150 neurons.

D. Applicability of the Model to Other Datasets

From Table V, we can see that for the data set in paper [17], the recognition rate of Model 4 with 150 neurons for bad photos is still acceptable as most of them can still be detected. The AUC value of Model 4 is greater than 0.5, so it has predictive value.

TABLE II. The Accuracy and Loss of Training and Validation Data

Model	Accuracy	Loss	Val Accuracy	Val loss
CNN	0.8020	0.4702	0.7160	0.5907
1	0.9480	0.1298	0.9140	0.2868
2	0.9945	0.0264	0.9300	0.1998
3	0.9980	0.0069	0.9560	0.1350
4	0.9980	0.0056	0.9560	0.1298
5	0.9500	0.1374	0.9320	0.2198

TABLE IV. RESULTS OF MODEL EVALUATION

Model	**Data set**	**Label**	**Confusion matrix**	**Precision**	**Recall**	**F1**	**ROC AUC**
1	Validation	Good 0	$\begin{vmatrix}471 & 29\\54 & 446\end{vmatrix}$	0.90	0.94	0.92	0.917
		Bad 1		0.94	0.89	0.91	
	Test	Good 0	$\begin{vmatrix}8 & 2\\2 & 6\end{vmatrix}$	0.80	0.80	0.80	0.775
		Bad 1		0.75	0.75	0.75	
2	Validation	Good 0	$\begin{vmatrix}471 & 29\\29 & 471\end{vmatrix}$	0.94	0.94	0.94	0.942
		Bad 1		0.95	0.94	0.94	
	Test	Good 0	$\begin{vmatrix}10 & 0\\1 & 7\end{vmatrix}$	0.91	1.00	0.95	0.938
		Bad 1		1.0	0.88	0.93	
3	Validation	Good 0	$\begin{vmatrix}488 & 12\\40 & 460\end{vmatrix}$	0.92	0.98	0.95	0.948
		Bad 1		0.97	0.92	0.95	
	Test	Good 0	$\begin{vmatrix}10 & 0\\1 & 7\end{vmatrix}$	0.91	1.0	0.95	0.938
		Bad 1		1.0	0.88	0.93	
4	Validation	Good 0	$\begin{vmatrix}477 & 32\\57 & 443\end{vmatrix}$	0.94	0.95	0.95	0.949
		Bad 1		0.95	0.94	0.95	
	Test	Good 0	$\begin{vmatrix}10 & 0\\1 & 7\end{vmatrix}$	0.91	1.0	0.95	0.938
		Bad 1		1.0	0.88	0.93	
5	Validation	Good 0	$\begin{vmatrix}475 & 25\\57 & 443\end{vmatrix}$	0.89	0.95	0.92	0.919
		Bad 1		0.95	0.89	0.92	
	Test	Good 0	$\begin{vmatrix}7 & 3\\2 & 6\end{vmatrix}$	0.78	0.70	0.74	0.725
		Bad 1		0.67	0.75	0.71	

TABLE V. SELECTION OF THE NUMBER OF NEURONS IN THE DENSE2 LAYER

Dense2	**Date sets**	**Label**	**Confusion matrix**	**Precision**	**Recall**	**F1**	**ROC AUC**
150 Model 4	Validation	0	$\begin{vmatrix}485 & 15\\27 & 473\end{vmatrix}$	0.95	0.97	0.96	0.958
		1		0.97	0.95	0.96	
	Test	0	$\begin{vmatrix}9 & 1\\0 & 8\end{vmatrix}$	1.0	0.9	0.95	0.950
		1		0.89	1.0	0.94	
256 Model 4	Validation	1	$\begin{vmatrix}480 & 20\\28 & 472\end{vmatrix}$	0.94	0.96	0.95	0.952
		0		0.96	0.94	0.95	
	Test	1	$\begin{vmatrix}10 & 0\\1 & 7\end{vmatrix}$	0.91	1.00	0.95	0.938
		0		1.0	0.88	0.93	

TABLE VI. RESULTS OF NEW TEST DATA BY BEST MODEL

Dense2	**Date set**	**Label**	**Confusion matrix**	**Precision**	**Recall**	**F1**	**ROC AUC**
150 Model 4	Test	1	$\begin{vmatrix}8 & 8\\4 & 10\end{vmatrix}$	0.67	0.50	0.57	0.607
		0		0.56	0.71	0.63	

VI. Discussion

1) Although the data sets of this paper were extracted from the conversation content of the other paper, it is clear that the keywords selected in Fig. 3 were similar to Fig. 4 (the distribution of paper [17]). At the same time, the training set and validation set are obtained by extracting keywords through TF-IDF technology, but accuracy of the test set (original photos) is very good. From both aspects proved that the TF-IDF technology is effective for expanding the data set from the content of the photos.

2) The selected positive impression photos were closely related to the daily life of the elderly, for example “famous attractions” (the elderly will travel with tour groups after retirement), “birthdays” (three generations celebrating their birthdays together), “zoos” (for visiting with their grandchildren). The negative impression pictures are all no longer related to their life, such as “toothpaste and toothbrush” (many elderly people wear dentures), “school life” (because many people did not have the opportunity to go to school in their youth), “sports” (they cannot maintain their physical strength) and so on. This phenomenon provides a general direction for us to choose photographs of conversations between two generations in the future.

3) To solve overfitting, the general approach is data augmentation [23]. As indicated on Table III in the CNN model, the accuracy of the validation set is almost 10% worse than the accuracy of the training set. The cause is that the expanded samples are still highly correlated. Therefore, for small data sets, it is not enough to consider only data amplification, and the model must be improved at the same time.

4) The deeper model VGG16 were chosen to transfer learning and fine-tuning , Table IV shown that the accuracy of the verification set and the training set similar and the accuracy of both increased to over 91%. The reason for this result is that models provided in this paper reduced the irrelevant information of photos. From the perspective of models, the main focus of preventing overfitting should be the entropy capacity of the model, or how much information the model can store [25, 27]. There are different ways to modulate entropic capacity, the main one is the choice of the number of parameters in your model, i.e. the number of layers and the size of each layer [35]. Although a model that can store more information may become more accurate by using more convolutions, there is a risk of overfitting to store content that is not related to classification features. Therefore more deeper model chosen with using transfer learning and fine-tuning to reduce the weights is necessary for small data sets.

5) Model 2, which fine-tuned the entire model, was superior to Model 1, which only trained the classifier. However, Model 2 was less accurate than Model 3 and Model 4. This was because Model 2 does not load the already trained classifiers from Model 1 like Model 3 and Model 4, but instead directly initializes a fully connected layer randomly on top of a pre-trained convolutional base layer. It leaded to large gradient updates triggered by the randomly initialized weights which wrecked the learned weights in the convolutional base.

Model 3, Model 4, and Model 5 are all loaded with the classifier weights already trained in Model 1. Among them, the accuracy of Model 3 and Model 4 is the same in the training set and the validation set, but Model 4 is better than Model 3 in the test set. Because Model 3's entire network had a very large entropic capacity and thus a strong tendency to overfitting.

The anti-overfitting effect of Model 5 is better than Model 3 and Model 4. The validation set is only 2% worse than the training set, but it performs the worst on the test set. Although the features learned by low-level convolutional blocks are more general and less abstract than those found higher-up, Model 5 only trains the last block (more specialized features) so is not enough for our data. Model 4 trained the last two blocks so is more effective for our data.

6) Related researches [1][16][17]has not verified their classification results with other data sets. But this paper used other data sets to verify its applicability to others. At the same time, [17] conversation experiment was done in Suzhou, [1] conversation experiment was done in Nanjing. This paper contains a questionnaire done in Shanghai. In the end, it was found to be inaccurate in recognizing photos in the good category, which shows that peoples' favorite things are usually highly personal. The model recognizes the bad category better. The first reason for this is because the photos with negative impressions are often something that has not been used recently by elderly people, such as toothpaste, toothbrushes, etc. Second, because the three locations tested belong to different cities in the south, it was easy to identify the differences in eating and cultural habits, such as not liking dumplings that are preferred by northerners. Therefore, in the future, we should avoid choosing photos that are not suitable for the conversation between the two generations from these two aspects.

VII. Conclusion and Perspectives

1) Using TF-IDF technology to extract keywords and then using a questionnaire to select photos provides is effective for expanding the data set based on the content of the photos.

2) Although Model 4 has a low recognition rate for selecting photos that are suitable for intergenerational conversation , it can effectively filter out photos that are not suitable for intergenerational conversation. Finally our goal is finished that this system filtered out photos that are unsuitable for conversation and kept the suitable photos.

3) This system was effective in the three cities of Suzhou, Shanghai and Nanjing, indicating that the system is likely to be applicable to more regions. Therefore, the next step is to select photos via the initially established photo dialogue support system, go to other cities for intergenerational dialogue, and then modify Model 4. So that this system can be applicable to most regions.

4) This system has some flaws, while select suitable photos for Intergenerational Conversation more factors should be considered, such as different age groups, different educational backgrounds, different hobbies, different personalities, etc. Later more models need be built to learn these characteristics.

ACKNOWLEDGMENT

This research was supported by JSPS KAKENHI grant number 15H01698. We would also like to express our gratitude to the China Scholarship Committee (CSC) and to all the young and the elderly who assisted in completing the experiment.

REFERENCES

[1] X. Dong, “Advances in the study of elderly care models in China,” Chinese Journal of Gerontology, vol. 039, no. 004, pp. 996-999, 2019.

[2] Y. D. Ya Zhu, Changqing Wang, “Research status and innovative thinking of healthy old-age care model,” Journal of Nanjing Medical University (Social Science Edition), vol. 018(002), pp. 103-106, 2018.

[3] S. He, “Research on the Mutual Assistance and Cooperative Social Pension Model from the Perspective of Developmental Welfare,” Rural Economy, no. 01, pp. 73-76, 2014.

[4] B. Li, and C. Jiang, “The Experience and Enlightenment of Community Pension Model in Britain and Japan,” Foreign trade, no. May 2015, pp. 58-59, 2015.

[5] Q. Wang, “Nursing service system and countermeasures for the elderly living alone and empty-nesters,” China Economic Times, vol. 12851, April 2019

[6] J. Liu, “A renew of research on the mental health of urban empty nest elderly in China,” Chinese Journal of Nursing, vol. 043, no. 5, pp. 457-459, 2008.

[7] Y. Xue, C. Zhang, H. Zhao, X. Zheng, and Y. Cai, “Depression status and influencing factors of empty nest elderly based on structural equation model,” Chinese Journal of Disease Control, vol. 23, no. 10, pp. 1181-1185, 2019.

[8] Z. Xiaochun, C. Dong-Eun, P. Siriaraya, and N. Kuwahara, “Sentiment Analysis and Classification of Photos for 2-Generation Conversation in China,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 10, 2019.

[9] A. Astell, N. Alm, G. Gowans, M. Ellis, R. Dye, J. Campbell, and P. Vaughan, “Working with people with dementia to develop technology: The CIRCA and Living in the Moment projects,” PSIGE newsletter, vol. 64, 2009.

[10] A. Astell, M. Ellis, N. Alm, R. Dye, J. Campbell, and G. Gowans, “Facilitating communication in dementia with multimedia technology,” Brain and Language, vol. 91, no. 1, pp. 80-81, 2004.

[11] A. J. Astell, N. Alm, G. Gowans, M. P. Ellis, R. Dye, and J. Campbell, “CIRCA: a communication prosthesis for dementia,” Technology and aging, pp. 67-76, 2008.

[12] A. J. Astell, M. P. Ellis, L. Bernardi, N. Alm, R. Dye, G. Gowans, and J. Campbell, “Using a touch screen computer to support relationships between people with dementia and caregivers,” Interacting with Computers, vol. 22, no. 4, pp. 267-275, 2010.

[13] N. Alm, R. Dye, G. Gowans, J. Campbell, A. Astell, and M. Ellis, “A communication support system for older people with dementia,” Computer, vol. 40, no. 5, pp. 35-41, 2007.

[14] B. A. Purves, A. Phinney, W. Hulko, G. Puurveen, and A. J. Astell, “Developing CIRCA-BC and exploring the role of the computer as a third participant in conversation,” American Journal of Alzheimer's Disease & Other Dementias®, vol. 30, no. 1, pp. 101-107, 2015.

[15] D. I. Fels, and A. J. Astell, “Storytelling as a model of conversation for people with dementia and caregivers,” American Journal of Alzheimer's Disease & Other Dementias®, vol. 26, no. 7, pp. 535-541, 2011.

[16] M. Iwamoto, N. Kuwahara, and K. Morimoto, “Comparison of Burden on Youth in Communicating with Elderly using Images Versus Photographs,” International Journal of Advanced Computer Science and Applications, vol. 6, no. 10, 2015.

[17] Z. Xiaochun, M. Iwamoto, and N. Kuwahara, “Evaluation of Photo Contents of Conversation Support System with Protocol Analysis Method,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 4, 2018.

[18] P. B. D. P. A. Vaidya, “Document clustering: TF-IDF approach,” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, pp. 61-66, 2016.

[19] A. A. H. A. E. K. I. E. M. Galinium, “Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach,” 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, pp. 1-4, 2014.

[20] A. Gulli, and S. Pal, Deep learning with Keras: Packt Publishing Ltd, 2017.

[21] K. B. Ahmed, Jelodar, Ahmad Babaeian, “Fine-Tuning VGG Neural Network For Fine-grained State Recognition of Food Images,” arXiv pre-print server, vol. abs/1809.09529, 2018.

[22] A. M. Dawud, K. Yurtkan, and H. Oztoprak, “Application of Deep Learning in Neuroradiology: Brain Haemorrhage Classification Using Transfer Learning,” Computational Intelligence and Neuroscience, vol. 2019, pp. 1-12, 2019.

[23] Y. Shima, “Image Augmentation for Object Image Classification Based On Combination of Pre-Trained CNN and SVM,” Journal of Physics: Conference Series, vol. 1004, pp. 012001, 2018.

[24] D.-X. Xue, R. Zhang, H. Feng, and Y.-L. Wang, “CNN-SVM for Microvascular Morphological Type Recognition with Data Augmentation,” Journal of Medical and Biological Engineering, vol. 36, no. 6, pp. 755-764, 2016.

[25] O. N. Belaid, and M. Loudini, “Classification of Brain Tumor by Combination of Pre-Trained VGG16 CNN,” Journal of Information Technology Management, vol. 12, no. 2, pp. 13-25, 2020.

[26] K. Rangasamy, M. A. As’ari, N. A. Rahmad, and N. F. Ghazali, “Hockey activity recognition using pre-trained deep learning model,” ICT Express, 2020.

[27] W. Setiawan, M. I. Utoyo, and R. Rulaningtyas, “Transfer learning with multiple pre-trained network for fundus classification,” Telkomnika, vol. 18, no. 3, 2020.

[28] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, “Breast cancer diagnosis with transfer learning and global pooling,” arXiv preprint arXiv:1909.11839, 2019.

[29] M. Lin, Q. Chen, and S. Yan, “Network in network,” arXiv preprint arXiv:1312.4400, 2013.

[30] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv preprint arXiv:1207.0580, 2012.

[31] A. Krizhevsky, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60(6), pp. 84-90, 2017.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[33] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, “Confusion Matrix-based Feature Selection,” MAICS, vol. 710, pp. 120-127, 2011.

[34] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." pp. 1015-1021.

[35] F. Chollet, “Building powerful image classification models using very little data,” Keras Blog, 2016.