Fashion Small Talk: Generating Friendly Comments on the Attire of the Interacting Person by using Image Captioning Technology*

Jaeyeon Lee, Dohyung Kim, Minsu Jang, Jaehong Kim

***Abstract*—People often starts conversation with small talks like little comments on the clothes, which is far harder for robots to make than talking about specific themes. Thus it would be helpful if a robot can make friendly comments on the attire of the interacting person in building emotional bond. We thought that image captioning technology can be utilized for the purpose. In this paper, we collected images of elderly people with variety of attires and asked the annotators to generate comments that she would make if she meets them in person. Then the image captioning network was trained with the collected training data. As a result, the trained network could make relatively sensible comments on the attire of the person in the image, which can be used for the robots to initiate a friendly conversation with the user.**

I. Introduction

Image captioning refers to a process that automatically generates natural language description of the given image [1]. It is considered as a part of a scene understanding, which has been a focus of extensive researches in computer vision society.

Typical image captioning approaches apply CNN (Convolutional Neural Network) to encode the image and RNN (Recurrent Neural Network) to decode the features into natural language description [2]. That is, when image and human generated annotation pairs are given, the network is trained to generate natural language description similar to the annotations on the given image.

The idea in this paper is that if the network can be trained to translate an image into a factual description, it would be also possible to generate opinions on the image. Of course, the human generated opinions should be given as training data in this case.

Thus, instead of collecting factual description on the image as in other image captioning DBs [1], we asked the annotators to generate friendly comments on the attire of the person in the image such as "You look young in vivid red shirts". We call these comments 'fashion small talk' that robots can use to initiate friendly conversation with humans.

II. Collection of Training Data

Because we are working on a project to develop service robot technologies mainly targeting on elderly people, we collected 6,062 images of elderly people with variety of attires from the internet (Fig. 1 shows an example of the images). Then we asked annotators to generate friendly comments on the attire the people in the image are wearing. As annotators, we recruited undergraduate students who major in fashion so that more sensible comments can be collected. There were 11 annotators and each generated one comment on an image, resulting in 6,062 x 11 = 66,682 comments (actually, the number was 66,445 because several were missing).

- **The floral design of the clothes looks unique.**
- **You look slim and elegant with that black top.**
- **The floral embroidery looks so good.**

Fig. 1 An Example of Collected Images along with Annotated Comments translated into English

As we can see in well-known image captioning DBs such as MSCOCO and flickr8K, the typical number of annotations per image is 5 [1]. However, unlike the factual descriptions which are more consistent because they almost always include major object or actions in the image, the comments can be more diverse according to the view of the annotator. Thus we thought richer annotations are needed for a reliable training. From the above 66,445 annotations, vocabulary of 14,556 words were extracted.

III. The Architecture of the Proposed Network

We used an open source image caption program as a basis and modified the program so that Korean language can be processed [3]. The network utilizes Inception V3 [4] for image feature extraction, which is followed by a fully connected layer to decrease the feature dimension to 300. On the other hand, the language model converts the word sequences into the embedding vectors and then passed through bidirectional LSTM.

Then the image model and language model are merged and passed through another bidirectional LSTM to output a vector

*This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00162, Development of Human-care Robot Technology for Aging Society)

Jaeyeon Lee, Dohyung Kim, Minsu Jang, Jaehong Kim are with the HRI Lab, ETRI, Daejeon, Republic of Korea ({leejy, dhkim008, minsu, jhkim504}@etri.re.kr).

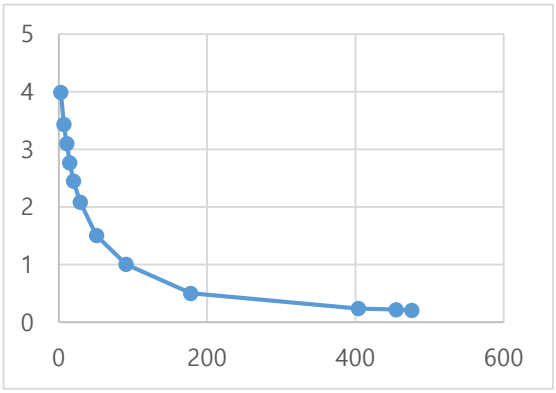

Fig. 2. Learning Curve

of vocabulary size. The highest probability value of the vocabulary vector is the next predicted word. This process goes on until the end of sentence mark comes out.

For the prediction, we tested 4 search methods, greedy search and beam searches of beam size of 3, 5, 7.

IV. Experiments

We divided the annotation data into 80% of training data and 20% of test data. Then the network is trained with batch size of 128 for 500 epochs. As shown in Fig. 2, the loss values decrease steadily until the loss value drops to 0.210. We saved the models whenever the new minimum loss was achieved, resulting in about 100+ models, from which we selected 12 models for the evaluation.

These 12 models and 4 search methods on 1,212 test images result in 58,176 generated comments, which is too much for human evaluation. Instead, we calculated BLEU score on the generated results [5]. The best BLEU-1 score was 0.33716 obtained from the model of epoch 7 – loss 4.045. This result shows that the smaller loss does not guarantee better result.

Several examples of the generated comments are shown in Fig. 3. The original comments are generated in Korean language and are translated into English for the readers. The generated comments are grammatically correct and generally make sense on the image. However, sometimes, the comments include wrong information such as referring to ‘red shirt’ while the person in the image wears ‘black jacket’.

V. Conclusion

People starts conversation with a small talk. If service robots can make interesting comments on the attire of the interacting user, it would be helpful in building emotional bond with the users.

In this paper, we tried to train image captioning network to generate friendly comments on the attire of the user instead of generating factual description of the scene. The result shows that the trained network could generate sensible comments in many cases, which can help the robot be more sociable.

References

[1] H. Wang, Y. Zhang, and X. Yu, "An Overview of Image Caption Generation Methods", Computational Intelligence and Neuroscience, Vol. 2020, Article ID 3062706.

[2] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), United Arab Emirates, 2019, pp. 246-251.

[3] F. E. Mustafa, “Keras Implementation of Image-Captioning”, https://github.com/Faizan-E-Mustafa/Image-Captioning.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2818-2826.

[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In ACL, 2002.

.

(a) The red clothes exhibit your unique personality.

(b) Neat and tidy style goes well on you.

(c) It looks good to wear when you go out.

Fig. 3. Examples of Generated Comments (Translated into English)