High-level Image Classification by Synergizing Image Captioning with BERT

Xiaohong Yu*, Yoseop Ahn*, and Jaehoon (Paul) Jeong*
* Department of Computer Science & Engineering, Sungkyunkwan University, Suwon, Republic of Korea
Email: {dawnyu, ahnjs124, pauljeong}@skku.edu

***Abstract*—Conventional image classification methods mostly aim to classify a single object in an image in which an object often occupies a large area. However, images in social network services (SNS) are more complicated. They usually include multiple objects that have much information, such as people, environments, and actions. In this work, we aim at understanding images from SNS and classifying them to categories such as fashion, traveling, education, beauty, and animals. To improve the classification accuracy in such complicated scenario, in this paper, we propose a new framework for high-level image classification by synergizing the image captioning and the Natural Language Processing (NLP) model. First, we use an image captioning model to understand images, which generates text descriptions about the images. Second, we use a natural language processing model to classify the generated text descriptions from the images. In this way, we can classify the images according to the classification results of the generated text descriptions. Our framework includes two models; one is image captioning model, which we use a TensorFlow based visual attention model with the inception V3 model for pre-processing and extracting the image features. The other model is the NLP model, Bidirectional Encoder Representations from Transformers (BERT). We have built a labeled image dataset from Instagram, a popular SNS platform, to test our framework. Our results show that our proposed method has a promising performance in terms of classification accuracy.**

***Index Terms*—High-level Image Classification, Image Captioning, BERT, SNS, COCO Dataset**

I. Introduction

Nowadays social network services (SNS), such as Instagram and Facebook, have attracted more and more people to use them. People share different information in SNS platforms, e.g., daily life, traveling experience, and delicious food, which bring great potential commercial opportunities to commercial companies. In order to advertise on Instagram, we need to look for appropriate influencers to maximize advertising effectiveness. Classifying influencers to different categories can help an advertisement client to fast find proper influencers with related interests. In Instagram platform, there are only images, videos, and post captions that can be analyzed for classification. Since most of the Instagram posts do not have any captions, to analyze these posts, we can only rely on the image and video information in these posts.

Image classification is a fundamental task that attempts to comprehend the entire information of an image as a whole. The goal of the image classification is to classify images by assigning them to different categories. Traditional image classification approaches categorize an image based on one major object appearing in the image. In another field, object detection approaches [1] involve both classification and object localization tasks, and is used to analyze more realistic cases in which multiple objects may exist in an image. However, for the images posted in SNS, both image classification and object detection techniques may still be far from fully understanding them.

In this paper, we propose a framework that uses an image captioning technique to generate text descriptions for images and a natural language processing (NLP) model to classify the generated text descriptions. A generated text description can include the number and the actions of objects as well as environment information in an image. Based on the generated text description of images, we use an NLP model, Bidirectional Encoder Representations from Transformers (BERT) [2], to analyze the image caption for a high-level image classification. To test our idea and method, we collected an image dataset with dimension 4×100, which has 4 labels and each label includes 100 samples. The main contributions of this paper are listed as follows:

- A high-level image classification framework that synergizes an image captioning model and an NLP model;
- Building an image data set with labels from Instagram for our framework;
- Extensive training and validation for the proposed framework.

The remainder of this paper is composed as follows: Section II summarizes the related work of image captioning and natural language classification technology. Section III describes our whole architecture and design of a high-level image classification model. The performance evaluation is presented in Section IV. In Section V, we conclude this paper along with future work.

II. Related Work

As the development of machine learning technology, many aspects of classification have greatly evolved, such as image classification, object detection [1], image captioning [3], [4] and dense captioning [5] in vision recognition area, and BERT [2] in NLP area. Usually, several datasets, e.g., MINIST [6], CIFAR-10 [7] and ImageNet [8], are used to check image classification performance, which are mostly focused on the handwritten character recognition and labeled single object tasks. Object detection techniques can recognize multiple objects in images, but they do not pay too much attention

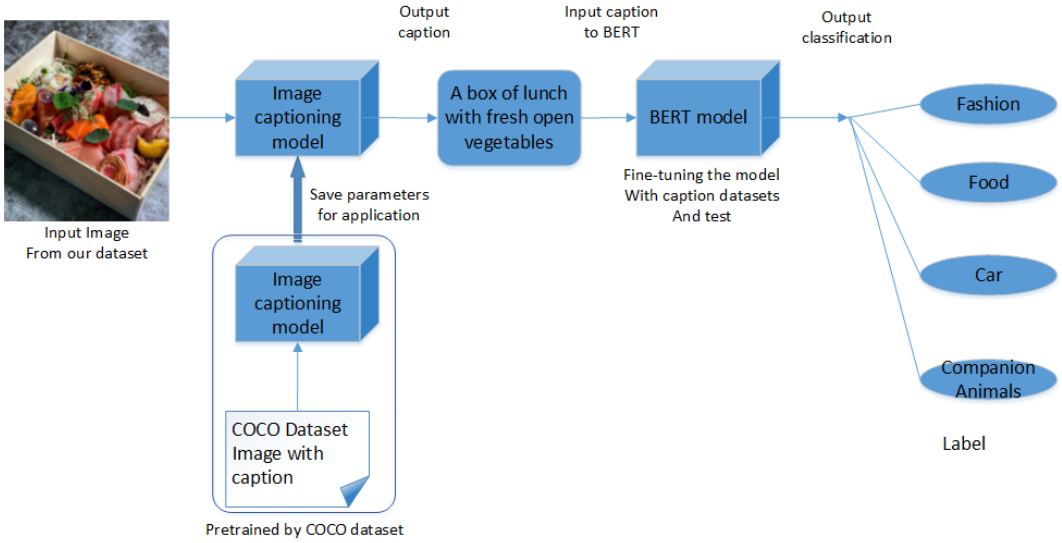

Fig. 1. High-level image classification architecture.

to the environment or background of an image, which makes them difficult to understand all the information in images. An image captioning technique attempts to recognize objects, object actions, and the background in an image, and then comprehends all the features about the image to generate sentences to describe the image.

Google firstly proposed the image caption generator in 2015 [3], extracted the images features using Vision Deep CNN encoder, and then made the features as inputs to the RNN decoder to generates sentences. Inspired by the attention technology, [4] introduced two attention-based image caption generators under a common framework. The encoder part is the same to the one used in [3], and only the decoder part is added with the attention mechanism while generating the next word, which means adding other location variables as the latent variables. [9] changed the final layer of CNN encoder and directly pass from image features to text.

For text analysis, BERT [2] is the state-of-the-art model in NLP area. BERT is a bidirectional transformer based pre-trained model developed by Google. It can be finetuned using different datasets for different downstream tasks, such as multi-label classification.

After converting an image to text using an image captioning model, we may understand the image better since more information is recorded by a generated text, which is good for our high-level image classification. In our project, there are many categories such as fashion, food, hot place and entertainment. Noted that it is difficult to categorize them by a traditional image classification technique.

III. Image Captioning and Classification

A. High-level Image Classification Architecture

Fig. 1 illustrates the overall architecture of our high-level image classification. There are two parts in our framework; the first part is an image captioning model that uses the Inception V3 model [10] to extract the image features and add the attention mechanism to generate image captions, and the second part is BERT model used for analyzing a caption and classifying it. There are several datasets for image captioning task, e.g., COCO dataset [11], TextCaps, and VizWiz. In this work, we use COCO dataset for pre-training of image captioning model, and save the checkpoint of the best performance for our framework. We generate image captions for our collected dataset, and use the captions to finetune the BERT model and test the classification performance.

B. Our Dataset

To test our idea and method, we collected 4×100 samples dataset, the images are all collected from Instagram using the Instagram API function. We collected four kinds of images, including fashion, food, car, and companion animals, which are also the labels for the images. The structure of our dataset is shown in Table I.

TABLE I
Dataset Structure

Media Link	Label
1.jpg	Fashion
2.jpg	Food
3.jpg	Car
4.jpg	Companion animals
....jpg	...

C. Image Captioning

COCO [11] is a large-scale object detection, segmentation, and captioning dataset. COCO has several features as follows: object segmentation, recognition in context, superpixel stuff segmentation, 1.5 million object instances, 80 object categories, and 5 captions per image. A sample of the COCO dataset is shown in Fig. 2.

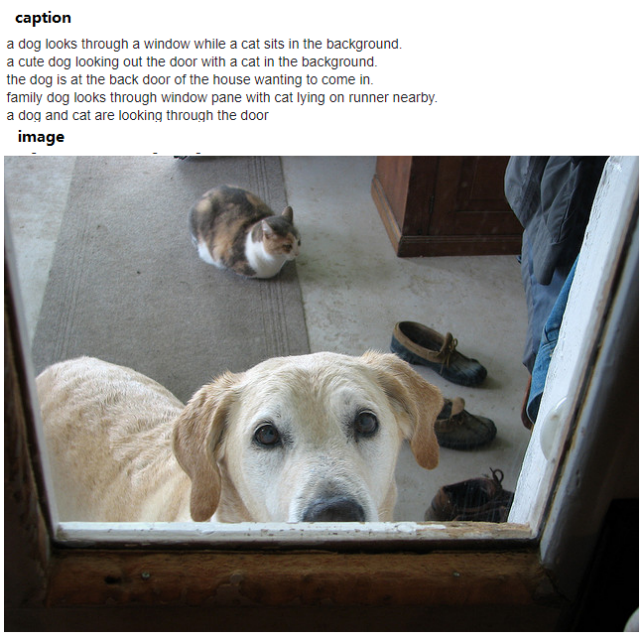

Fig. 2. COCO dataset structure.

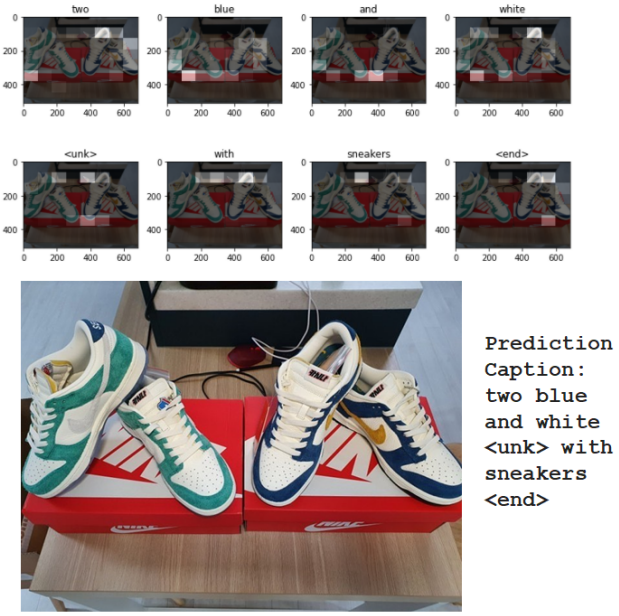

Fig. 3. An output of image captioning part.

'caption'	'link'	'label'
a young person standin	https://scontent-ssn1-	0
a high on the ground w	https://scontent-ssn1-	0
a boy with a man stand	https://scontent-ssn1-	0
a woman is in front of a	https://scontent-ssn1-	0
a young woman in a <u	https://scontent-ssn1-	0
a woman being <unk> p	https://scontent-ssn1-	0
a boy is taking a purple	https://scontent-ssn1-	0
a woman smiling in the	https://scontent-ssn1-	0
the young woman stan	https://scontent-ssn1-	0
a person wearing glasse	https://scontent-ssn1-	0
a cookie on a white liqu	https://scontent-ssn1-	0
a woman in a white t sh	https://scontent-ssn1-	0
a smiling girl with black	https://scontent-ssn1-	0

Fig. 4. An example of dataset for the generated image captions.

Inception V3 model is a widely used image recognition model that has shown a greater-than 78.1% accuracy on the ImageNet dataset. In this paper, we use the Inception V3 model to preprocess and extract the image features from the collected images.

The visual attention refers to a set of cognitive operations that mediate the selection of relevance and filter out irrelevant information from cluttered visual scenes. Here, we use the visual attention mechanism to check which part of the image the model are focusing on as it generates a caption.

To generate a proper caption for an image, first, we need to train the Inception V3 model using the COCO dataset to extract image features. Second, we use the decoder-encoder structure of an attention technology to locate the attention of the image and predict the next word. Finally, we can predict the caption of images and check the performance using a test subset of the COCO dataset by saving the checkpoint of whole model. The model with a saved checkpoint is used for our dataset to generate captions. One output of image captioning model is shown in Fig. 3. Fig. 4 shows an example of the image captions generated from the COCO dataset.

D. BERT and Classification

BERT is a state-of-the-art pretrained model working on a variety of NLP tasks. In this paper, we use the base version of BERT (i.e., bert-base-uncased). It has 12 layers, 12 attention heads, and 110 million parameters. As our dataset has 4 labels, we need to define the BERT classifier according to our classes.

From image captioning part, we can generate the new dataset with caption for each image, and then use the image captions and labels to fine-tune the BERT model, and eventually check the performance. The new dataset is shown in Table II.

TABLE II
NEW GENERATED DATASET FORMAT

Media Link	Caption	Label
1.jpg	A woman wearing a dress	Fashion
2.jpg	A bow of noodles with fresh vegetable	Food
3.jpg	A silver SUV is parking	Car
4.jpg	A white dog is perched in the room	Companion animals
....jpg	...	...

IV. PERFORMANCE EVALUATION

The implementation of our framework is carried out in Google Colab notebook [12]. The proposed framework con-

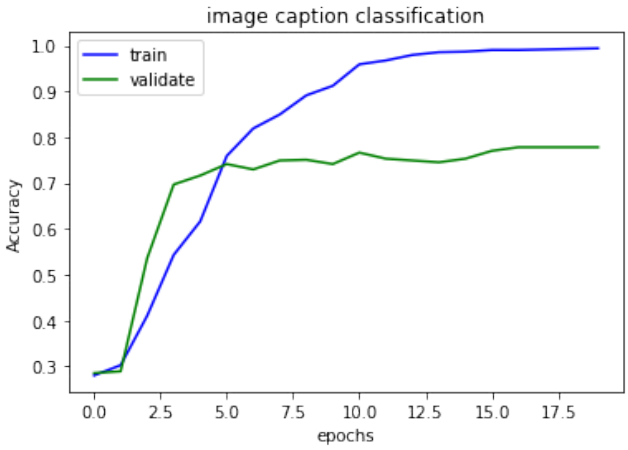

Fig. 5. Classification accuracy in training and validation for image caption.

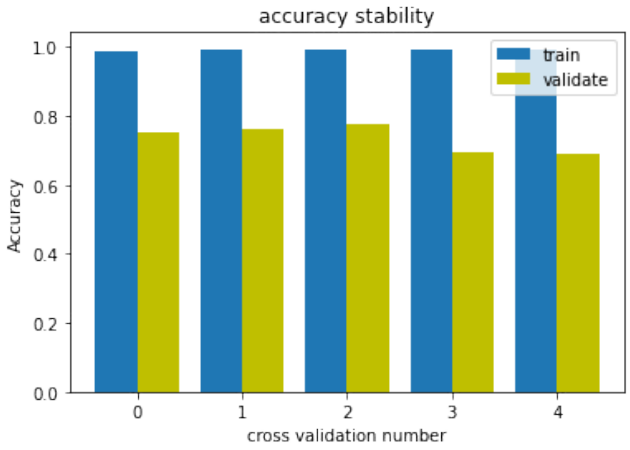

Fig. 6. Stability of classification accuracy for both training and validation stages.

sists of two parts, the first part is for image captioning and the second part is caption classification using the BERT model.

A. Classification Accuracy

The dataset of the generated image captions is divided into two groups, such as training group and validation group. We first use the training group to fine-tune the BERT model and then use the validation group to test the actual accuracy of our trained model. Fig. 5 shows the classification accuracy in training stage in the proposed architecture. We trained the BERT model by the generated captioning dataset with 20 epochs. As we can see the blue line in Fig. 5, the classification accuracy of the image caption in the training stage was continuously growing. When the epoch reaches about 15, the accuracy is close to 1. This shows that our training model is correctly classifying the captions.

When training the model, we use the validation group to check the actual accuracy of our trained model in order to look for the model with good performance. The green line in Fig. 5 shows the classification accuracy in the validation stage. As it shows, the classification accuracy in this validation stage after about 3 epochs is a little higher than 0.7. In the fifth epoch, the training accuracy equals the validation accuracy. After that, the training accuracy increases, but the validation accuracy becomes stable. The accuracy in validation becomes much lower than the accuracy in training. This is because that the trained model overfits the dataset of the training group. When the model is applied to another group of the dataset, the accuracy naturally decreases. Another impact factor is the quality of generated captions in the image captioning part. If the generated captions for the image dataset cannot correctly reflect the major information shown in those images, the accuracy of this caption classification would also be reduced.

B. Classification Stability

To show the stability of the proposed classification architecture, we also conducted cross validation. Different from the previous training process, this cross validation partitions the dataset into several small groups where each group has both training and validation parts. As shown in Fig. 6, we conducted this cross validation 5 times and each time has a training accuracy and validation accuracy. For the training part, the classification accuracy is stable at about 0.98. However, the accuracy in validation is reduced down to about 0.75. Although the accuracy in the validation part becomes low, the average accuracy shows a stable performance, which demonstrates that the proposed system can steadfastly classify images based on the generated captions. Another reason for the test accuracy being low is that in image captioning part, we pretrained the model using the COCO dataset, in which the image information may have much difference with our dataset from Instagram. This may cause that the caption generated for our dataset has the style of COCO dataset caption.

V. Conclusion

In this paper, we provide a new method for high-level image classification. The image captioning model with an attention architecture can better understand the image features and generate proper captions. The BERT model for caption classification is also very flexible to various kinds of text input. In this paper, we only selected four classes for labeling images in verifying our idea. For the future work, we plan to optimize the image captioning model to generate a more intelligent and accurate captions for images, which shall bring a better performance to our proposed high-level image classification system. We will also optimize our dataset with more categories.

Acknowledgments

This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2017-0-01633) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was supported in part by the IITP grant funded by the MSIT

(No. 2019-0-01343, Regional strategic industry convergence security core talent training business). Note that Jaehoon (Paul) Jeong is the corresponding author.

REFERENCES

[1] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, p. 128837–128868, 2019. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2939201

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2016.

[5] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," 2015.

[6] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[7] A. Krizhevsky, "Learning multiple layers of features from tiny images," *University of Toronto*, 05 2012.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[9] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" 2016.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

[11] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 2015.

[12] Google, "Colaboratory, Google Research," Accessed on 05/31/2021. [Online]. Available: https://colab.research.google.com/