Image Captioners Sometimes Tell More Than Images They See

Honori Udo*   Takafumi Koshinaka

The School of Data Science, Yokohama City University

22-2 Seto, Kanazawa-ku, Yokohama 236-0027 Japan

* Now with NTT Comware Corporation

Abstract

Image captioning, a.k.a. "image-to-text," which generates descriptive text from given images, has been rapidly developing throughout the era of deep learning. To what extent is the information in the original image preserved in the descriptive text generated by an image captioner? To answer that question, we have performed experiments involving the classification of images from descriptive text alone, without referring to the images at all, and compared results with those from standard image-based classifiers. We have evaluate several image captioning models with respect to a disaster image classification task, CrisisNLP, and show that descriptive text classifiers can sometimes achieve higher accuracy than standard image-based classifiers. Further, we show that fusing an image-based classifier with a descriptive text classifier can provide improvement in accuracy.

1. Introduction

Advances in neural network-based representation learning have enabled "embedding," which maps any type of data into a latent space with a certain number of dimensions. Numerous methodologies have been proposed for handling media data with different modalities [10, 12]. In particular, significant research results have recently been reported in the field dealing with images and texts referred to as "Vision and Language" (V&L) [3]. Typical tasks here would include visual question answering (VQA) for answering questions with images [4], image captioning for assigning descriptions to images [21], and image generation for generating images from descriptions [11].

Image captioning (image-to-text), which is addressed in this paper, was inspired by sequence-to-sequence learning in neural machine translation [18] and has shown remarkable progress employing a similar approach, one in which a context vector obtained by encoding an input image is decoded to generate descriptive text in an auto-regressive manner [22]. It has a two-sided relationship with image generation (text-to-image), which has attracted public attention with the advent of Dall-E 2 and subsequent approaches to generation [14, 15]. Further development is expected with the support of such foundation models for Vision and Language as CLIP [13] and BLIP [8], which are trained with a large number of images and large amount of text found on the Internet.

Image captioning has something in common with automatic speech recognition (ASR a.k.a. speech-to-text), which has long received much research attention. While ASR converts acoustic signals caused by air vibration into text, image captioning converts light signals into text. ASR extracts only linguistic information from input speech and discards such non-linguistic information as tone, emotion, and the gender and age of the speaker. In this sense, ASR can be viewed as a kind of feature extraction. Image captioning is similar in that it extracts certain information from an input image and discards other information. We aim to clarify what that "certain information" actually is.

In speech emotion recognition, Srinivasan et al. [17] employ linguistic information (text) obtained from ASR as features and show that the emotion-recognition accuracy is superior to that with conventional methods that use only acoustic features. This may also apply in image captioning. Could an image captioner also serve as a feature extractor that complements image recognition? In this paper, we examine how accurately image classification can be performed using descriptive text obtained by an image captioner as a feature. We also try to improve classification accuracy by combining it with standard image classifiers based on convolutional neural networks (CNNs). Our experimental setup is designed for use with CrisisNLP [2, 1], a benchmark dataset for disaster image classification. There have been earlier studies that attempted multi-task learning with image classification and captioning as two parallel tasks [24], but, to the best of our knowledge, this is the first attempt to use an image captioner as a feature extractor for image classification.
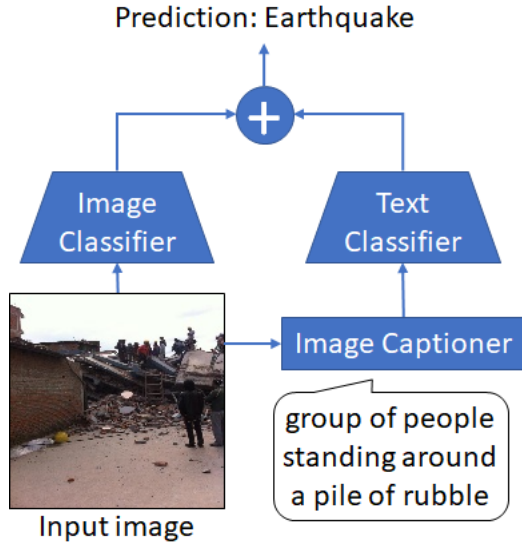
Figure 1. System configuration: a standard image-based classifier (left half) and a text-based classifier combined with an image captioner (right half).

The remainder of this paper is organized as follows. Section 2 describes the configuration of our image classification system using image/text-based classifiers combined with an image captioner. The experimental setup and results for image- and text-based single-modal systems as well as fused multi-modal systems are presented in Section 3. Section 4 summarizes our work.

2 System Configuration

Figure 1 shows the configuration of the final form of the image classification system considered in this paper. The left half shows a standard image classifier for use with a neural network such as a CNN. The right half shows the connection of an image captioner with a text classifier, in tandem, for the classification of images on the basis of linguistic information extracted from images.

2.1 Image and Text Classifiers

We used pre-trained models throughout. MobileNetV2 [16] and EfficientNet [20] were used for the image-based classifier, and BERT$_{BASE}$ [6] was used for the text-based classifier after fine-tuning with the training data for the target task. Regarding hyper-parameter settings in fine-tuning, such as learning rate, mini-batch size, number

of epochs, data augmentation, dropout, etc., we follow publicly available tutorial codes [1][2][3].

2.2 Image Captioners

We focused on three image captioning models. None of them were fine-tuned using the data for the target task (because no image description text was available for the target task), and the original models were used as it is.

InceptionV3+RNN: This is a basic model of small scale that encodes an input image into a vector by using InceptionV3 [19] and then decodes it to generate a caption using a recurrent neural network (Gated Recurrent Unit; GRU [5]). An attention mechanism [23] is equipped between the encoder and decoder, and features of each part of the image are selectively sent to the decoder. The entire system was trained with the MS-COCO dataset [9]. We followed TensorFlow's tutorial implementation [4].

BLIP: This is a foundation model that has learned a large number of images and amount of text and is applicable to a wide range of Vision and Language tasks [8]. When used as an image captioner, it takes the form of an encoder-decoder configuration based on the Transformer model. Users can easily run the sample code (demo.ipynb) on GitHub [5] to obtain captions for their own images. BLIP is a relatively advanced, large-scale model that is capable of producing quite accurate captions.

CLIP Interrogator: Given an image, this model infers prompts for such AI image generators as Stable Diffusion and Midjourney so as to generate similar images. Since the text generated by CLIP Interrogator is not meant to be read by humans, it may not be considered an image captioner in the strict sense, but we tested it as a model that can generate richer text than can BLIP.

Although the technical specifications of CLIP Interrogator have not been published as a paper and there is no literature that can be referred to, it may be presumed from the code [6] and its operation that it first generates a base caption using BLIP and then selects and adds phrases that match the target image from a predefined set of phrases called *Flavors*. CLIP image/text encoders [13] are used to measure the degree of matching between a target image and the phrases

---

[1] Transfer learning and fine-tuning `https://www.tensorflow.org/tutorials/images/transfer_learning`
[2] Image classification via fine-tuning with EfficientNet `https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/`
[3] Classify text with BERT `https://www.tensorflow.org/text/tutorials/classify_text_with_bert`
[4] Image captioning with visual attention `https://www.tensorflow.org/tutorials/text/image_captioning`
The architecture of the decoder is Embed(256)–GRU(512)–FC(512)–FC.
[5] `https://github.com/salesforce/BLIP`
[6] `https://github.com/pharmapsychotic/clip-interrogator`

in *Flavors*. Flavors contains approximately 100,000 words and phrases, including those referring to objects and entities (e.g., motorcycle, building, young woman), image styles (e.g., photo-realistic), and artist names (e.g., greg rutkowski). We used the code released by the developer (clip_interrogator.ipynb, version 2.2).

It should be noted that the default settings are different for the standalone BLIP and the BLIP running within the CLIP Interrogator, and that the latter emphasizes the quality of captions by changing search parameters (e.g., num_beams). In our experiment, in order to match their operations, the parameters of the standalone BLIP were matched with those of the CLIP Interrogator.

2.3 System Fusion

As previously indicated in Figure 1, we fuse the outputs of an image-based classifier with those of a text-based classifier to improve classification accuracy. Fusion methods include feature-level fusion (early fusion), which inputs the hidden layer states of each classifier into another neural-network classifier, and score-level fusion (late fusion), which averages the classification results of each classifier. Here we choose the latter for simplicity. That is, if the number of classes is C and the output of the image/text-based classifiers normalized by the softmax function are $\boldsymbol{y}^{(I)} = \left(y_1^{(I)}, \cdots, y_C^{(I)} \right)$ and $\boldsymbol{y}^{(T)} = \left(y_1^{(T)}, \cdots, y_C^{(T)} \right)$, respectively, then the classification result obtained with score-level fusion may be calculated as $\boldsymbol{y} = (1-w)\,\boldsymbol{y}^{(I)} + w\boldsymbol{y}^{(T)}$, where $0 \leq w \leq 1$ is the weight coefficient for the text-based classifier.

Table 1. Number of images and classes contained in two tasks defined in the CrisisNLP dataset.

Task	Disaster types	Damage severity
Train	12,724	26,898
Dev	1,574	2,898
Test	3,213	5,100
Num classes	7 (earthquake, fire, flood, hurricane, landslide, other disaster, not disaster)	3 (severe damage, mild damage, little or none)

Figure 2. Example images for different disaster types included in the CrisisNLP dataset (cited from [2]). There is another type referred to as "not disaster," which is not shown here.

3 Experiments

We used the CrisisNLP dataset [2], which is a collection of natural disaster images shared on such social media as Twitter [7]. The dataset provides four image classification tasks, for each of which training (Train), development (Dev), and test (Test) data partitions are defined. Among them, we focus on two tasks: 1) "Disaster types," for predicting types of disasters, such as earthquake, flood, etc.; 2) "Damage severity," for predicting the degree of damage caused by a disaster in terms of three stages. (See Table 1 and Figure 2)

We first show the classification accuracies of single-modal systems using only an image-based classifier or a text-based classifier (Table 2). To reduce the randomness of model parameter initialization in fine-tuning, each of those accuracies is averaged over five trials.

Regarding the image-based classifiers, EfficientNet models of three sizes (B0, B1, B2) outperformed MobileNetV2 while no significant difference was observed among the three. Comparing the text-based classifier with three different image captioners, we can first see that the most basic image captioner, InceptionV3+RNN, falls far short of the classification accuracy of standard image-based

[7]https://crisisnlp.qcri.org/crisis-image-datasets-asonam20

Table 2. Accuracies (%) of image-based classifiers: MobileNetV2 (MobNetV2), EfficientNet (EffNet) and text-based classifiers: BERT combined with InceptionV3+RNN (IV3+RNN), BLIP (BLIP), CLIP-I (CLIP Interrogator).

System	Disaster types	Damage severity
Image-based		
MobNetV2	69.79	73.37
EffNet (B0)	**76.48**	**76.84**
EffNet (B1)	75.05	76.71
EffNet (B2)	76.33	76.31
Text-based		
IV3+RNN	42.40	52.36
BLIP	71.14	72.17
CLIP-I	**85.28**	**78.67**

classifiers. A look at the captions generated by InceptionV3+RNN reveals that most of them are seemingly irrelevant with respect to the images, and it seems difficult to predict either the type of disaster or the degree of damage from these captions (as indicated later in Figure 3). By way of contrast, BLIP did generate good captions for many images. The caption in previously shown in Figure 1 is one actually generated by BLIP, and it notes such important elements in the image as "people" and "rubble." The text-based classifier with BLIP consequently achieved much better accuracy, roughly comparable to that of standard image-based classifiers. The accuracy of the text-based classifier using CLIP Interrogator (CLIP-I) is even higher, and results significantly exceed those of standard image-based classifiers. These results suggest that foundation models of Vision and Language trained on a large amount of image/text data can be an effective feature extractor for image classification.

Figure 3 shows an example of image classification results for an image that should be classified as "earthquake." As previously noted, InceptionV3+RNN (IV3+RNN), which was the most basic image captioner, produced a description that was far from the actual content of the image. BLIP's descriptions were generally much more accurate. The CLIP Interrogator (CLIP-I) behaved quite differently from these two. After beginning with a common sentence from BLIP, it continued the explanation with a large number of phrases selected by CLIP. Although those phrases make no sense as sentences, we can observe some that reflect the true class, such as "earthquake" and "collapsed building." On the other hand, there are also such completely irrelevant phrases as "akira movie style" and "videogame still;" they would, however, be good clues for Stable Diffusion to use to generate images.


Input	
MobNetV2	→ **not disaster**
EffNet B0	→ **earthquake**
IV3+RNN	a restaurant that are shining that is lined up with lots of people. → **not disaster**
BLIP	a group of people standing on top of a building → **not disaster**
CLIP-I	a group of people standing on top of a building, collapsed building, buildings collapsed, collapsed buildings, videogame still, burning building, building destroyed, background of resident evil game, 19xx :2 akira movie style : 8, photo", big impact hit on the building, damaged buildings, earthquake, unreal engine. film still, test, building on fire → **earthquake**

**Figure 3. Example results with an image to be classified as "earthquake."**

Figure 4 shows the results of score-level fusion that averaged the output of the image/text-based classifiers with weight w, where MobileNetV2 was used for the image-based classifier. Both of the two tasks (Disaster types, Damage severity) show similar trends, i.e., when using sufficiently good models, such as BLIP and CLIP Interrogator (CLIP-I), for image captioning, the classification accuracy can be improved by appropriately choosing the weight w. It seems that image captioning models extract features different from CNN models; in other words, they look at images from a different perspective than do CNNs. Unfortunately, when using InceptionV3+RNN for image captioning, the effect of fusion was nealy unseen because the difference in performance between the two modalities was too large.

(a) Disaster types
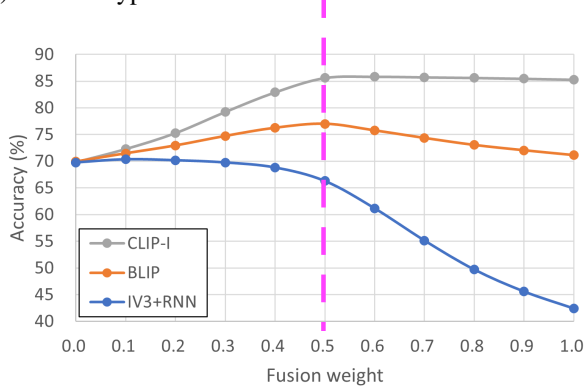
(b) Damage severity
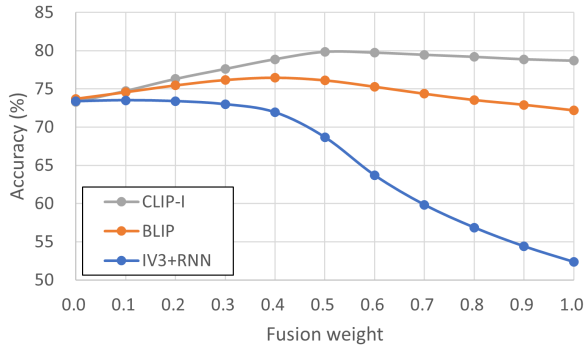
Figure 4. Score-level fusion results using Mo-bileNetV2 as the image-based classifier: Horizontal axis stands for the fusion weight w for the text-based classifier. $w = 0$ **and** $w = 1$ **correspond to image-based and text-based single-modal systems, respectively.**

When we fused the text-based classifier with another image-based classifier, EfficientNet (B0), we could still see a synergistic effect, as indicated in Figure 5. We should note that, as the classification accuracy of EfficientNet is higher than that of the text-based classifier using BLIP (un-like MobileNetV2), the synergistic effect is not so large as in the case of MobileNetV2.
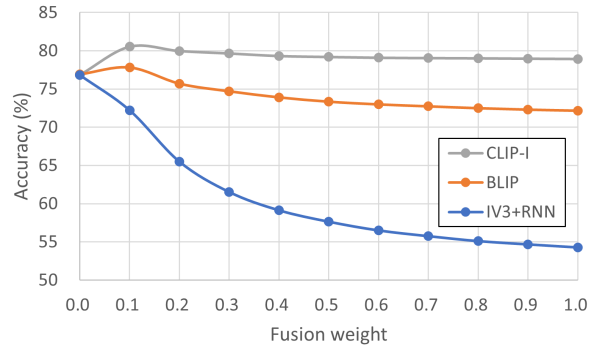
4 Summary

Starting from the question of what kind of information image captioners can extract from images, we performed image classification using only linguistic information contained in captions, and have shown that it is possible to achieve classification accuracy that surpasses that of stan-

(a) Disaster types
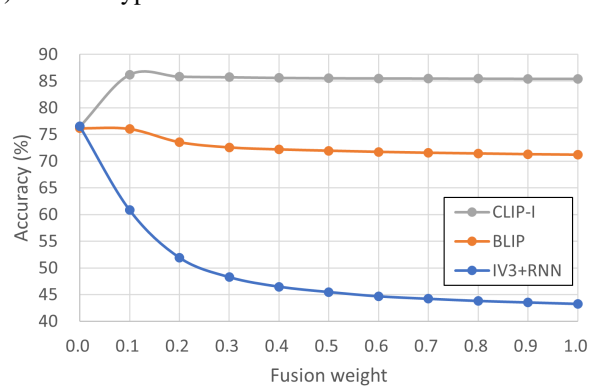
(b) Damage severity

Figure 5. Score-level fusion results using Ef-ficientNet B0 (rather than MobileNetV2 in Figure 4) as the image-based classifier.

dard image-based classifiers using CNNs. Further, we have confirmed that synergistic effects can be obtained by fusion with those image-based classifiers. It can be said that image captioners based on large-scale foundation models are effective feature extractors for image classification. Image captioning is a rapidly evolving area, and we need to keep up with new technologies that are being released one after another [7] [8].

In this study, we experimented with a simple system configuration, in which very basic parts had been combined. In future work, we intend to introduce more advanced methods, such as feature-level fusion and knowledge distillation [17], in order to search for better answers to the original question. It should be further noted that the behavior of image captioners would seem to depend heavily on the images to be captioned. While scenes of disasters are rela-

---

[8]https://docs.midjourney.com/docs/describe

tively easy to describe, the characteristics of a human face would not be so easy. Studies on other datasets are to be included in our future work.

Acknowledgment

References

[1] F. Alam, F. Ofli, M. Imran, T. Alam, and U. Qazi. Crisis-MMD: Multimodal twitter datasets from natural disasters. In *12th International AAAI Conference on Web and Social Media (ICWSM)*, 2018.

[2] F. Alam, F. Ofli, M. Imran, T. Alam, and U. Qazi. Deep learning benchmarks and datasets for social media image classification for disaster response. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In *36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In *IEEE Int'l Conf. on Computer Vision (ICCV)*, 2015.

[5] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[7] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023.

[8] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv:2201.12086*, 2022.

[9] 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. In *European Conference on Computer Vision (ECCV)*, 2014.

[10] M. R. Makiuchi, K. Uto, and K. Shinoda. Multimodal emotion recognition with high-level speech and text features. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

[11] E. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov. Generating images from captions with attention. In *International Conference on Learning Representations (ICLR)*, 2016.

[12] C. Narisetty, E. Tsunoo, X. Chang, Y. Kashiwagi, M. Hentschel, and S. Watanabe. Joint speech recognition and audio captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021.

[14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

[15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] S. Srinivasan, Z. Huang, and K. Kirchhoff. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[18] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence-to-sequence learning with neural networks. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] M. Tan and Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

[21] Y. Ushiku, Y. Mukuta, M. Yamaguchi, and T. Harada. Common subspace for model and similarity: Phrase learning for sentence generation from images. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.

[24] Y. C. Yoon, S. Y. Park, S. M. Park, and H. Lim. Image classification and captioning model considering a cam-based disagreement loss. *ETRI Journal*, 42(1), 2020.