JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 711-724 (2023)
DOI: 10.6688/JISE.202307_39(4).0001

XAI for Image Captioning using SHAP

CHRISTINE DEWI[1], RUNG-CHING CHEN[2,+], HUI YU[3] AND XIAOYI JIANG[4]
[1]*Department of Information Technology*
Satya Wacana Christian University
Salatiga 50517, Indonesia
[2]*Department of Information Management*
Chaoyang University of Technology
Taichung City, 413310 Taiwan
[3]*School of Creative Technologies*
University of Portsmouth, PO1 2UP UK
[4]*Department of Mathematics and Computer Science*
University of Münster
D-48149 Münster, Germany
E-mail: crching@cyut.edu.tw

In the fields of computer vision (CV) and natural language processing (NLP), they attempt to create a textual description of a given image is known as image captioning. Captioning is the process of creating an explanation for an image. Recognizing the significant items in an image, their qualities, and their connections are required for image captioning. It must also be able to construct phrases that are valid in both syntax and semantics. Deep-learning-based approaches are deal with the intricacies and problems of image captioning. This article provides a simple and effective Explainable Artificial Intelligence (XAI) technique for image text. Deep learning techniques have been widely applied to this work in recent years, and the results have been relatively positive. This work employs Azure Cognitive Service and Open-Source Image Captioning model to get image caption. We implement Explainable Artificial Intelligence (XAI) Image Captioning (Image to Text) using Shapley Additive explanations (SHAP). This work applies Cosine similarity by spaCy and Term Frequency Inverse Document Frequency (TF-IDF transform) to evaluate the sentence similarity. Our research work found that Azure Cognitive Services provides better descriptions for images compared to the Open-Source Image Captioning Model.

Keywords: SHAP, explainable artificial intelligence, image captioning, azure cognitive service, API

1. INTRODUCTION

Image captioning automatically creates natural language descriptions for images [1, 2]. Additionally, image captioning is helpful for a variety of purposes, including image retrieval [3, 4], assisting the visually impaired [5, 6], and Intelligence human-computer interaction. It has been a difficult cross-disciplinary project for decades, requiring both computer vision and natural language processing.

Recently, research on pictorial text or image presentation with meaningful phrases using advanced deep learning methods and computer vision ideas has grown significantly. Image captioning is frequently used in various situations, such as to assist blind people by converting text to speech with real-time output [7, 8]. As important as the ac-

Received January 26, 2022; revised February 27, 2022; accepted April 20, 2022.
Communicated by Mu-Yen Chen.
[+] Corresponding author.

curacy of predictions, explainable AI (XAI) technology can explain why machine learning (ML) models make certain predictions. This technology can explain why machine learning (ML) models make certain predictions [9, 10]. Aside from that, XAI provides a chance to make the decision-making process more transparent and efficient [11]. Designing Intelligence systems that can explain their predictions or recommendations to humans is the goal of XAI research [12]. XAI approaches enable blind and visually impaired (BVI) people to carry out their primary activities with little or no assistance from others, reducing their reliance on others [13, 14]. A variety of audio devices are available to BVI molecular scientists to assist them in reading text in articles and working with computers [15]. We implement XAI Image Captioning (Image to Text) in our experiment using Shapley Additive explanations (SHAP). Microsoft Azure Cognitive Services is a collection of APIs and SDKs that enable developers to build Intelligence applications by making artificial intelligence capabilities available to everyone, including those without machine-learning expertise. Azure Cognitive Service is being used in our experiment to get image captions. After that, we will compare the result with Open-Source Image Captioning Model. Comparing sentences is indicated by the degree of probability that the sentences are related to each other. Furthermore, detecting sentence similarity is an important problem in various applications. Further, we evaluate the similarity of sentences in image caption employing Cosine similarity by spaCy and Term Frequency Inverse Document Frequency (TF-IDF transform).

Briefly summarized, the most important contributions made by this research are as follows: (1) XAI Image Captioning (Image to Text) using Shapley Additive explanations (SHAP); (2) Employ Azure Cognitive Service and Open-Source Image Captioning model to get image caption; (3) We analyzed and discussed in detail the experiment result; (4) Azure Cognitive Services performs well compared to other methods based on the experimental results; (5) Our research work employs Cosine similarity by spaCy and Term Frequency Inverse Document Frequency (TF-IDF transform) to evaluate the sentence similarity. The remainder of this paper is structured in the following manner. The following sections, Sections 2 and 3 provide information on related works and our methodology, respectively. Section 4 discusses the research findings and results. Finally, the conclusion and future works are described in Section 5.

2. MATERIALS AND METHODS

2.1 Image Captioning and Explainable Artificial Intelligence (XAI)

Image Captioning is critical for a variety of reasons. It may be used, for example, to do automated picture indexing. Because picture indexing is critical for content-based image retrieval (CBIR) [16], it applies to a wide variety of fields, including biology, commerce, the military, education, digital libraries, and online search. The research on image captioning can be categorized into three classes: (1) Template-based approaches [17]; (2) Retrieval-based approaches [18]; (3) Generation-based approaches. According to Karpathy *et al.*, a deep fragment embedding approach is proposed to match image-caption pairs based on the alignment of visual segments (the detected objects) and caption segments, which include subjects, objects, and verbs, in order to improve matching accuracy. Mao *et al.* [19] proposed a multimodal recurrent neural network (m-RNN)

method for generating novel image captions. This method has two subnetworks: a deep recurrent neural network for sentences and a deep convolutional image network [20]. We use 60 images from ImageNet [21] to conduct our experiment in our work. Fig. 1 displays the ImageNet dataset example that we use in our experiment. Most of the image sizes are 224×224 pixels, and we choose 60 images randomly from the ImageNet.

Fig. 1. ImageNet dataset example.

Artificial Intelligence (AI) capabilities have improved significantly in recent years, but contemporary methods such as deep neural networks are becoming more complex resembling black boxes. This raises the issue of how reliable AI forecasts are and the critical factors in achieving widespread acceptance in society and business. As a result, technologies that address these issues and enable explainable artificial intelligence (XAI) are in high demand [22]. Han and Choi [23] have proposed an explainable image captioning model, which provides a visual link between the region of an object in the given image and the particular word or phrase in the generated sentence. Because causal links between features can be defined directly using graph structures, Holzinger A *et al.* [24] emphasizes that Graph Neural Networks play a significant role in the analysis. Using their research, they hope to inspire the international XAI community to continue its research into multimodal embedding and interactive explain ability in order to lay the groundwork for future human-AI interfaces that are effective.

2.2 Shapley Additive Explanations (SHAP)

SHAP is the state-of-the-art Machine Learning explain ability, and it is available for free. Developed by Lundberg and Lee in 2017 [25], this method provides a great approach to reverse-engineer the output of any prediction algorithm. Fig. 2 depicts a high-level overview of how to understand the predictions of any model using the SHAP algorithm. The SHAP value provides two critical advantages as follows:

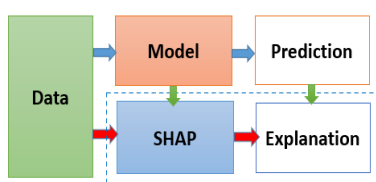


Fig. 2. SHAP overview to explain the model.

(1) The SHAP value may be computed for any model, not only simple linear models, for a variety of reasons; (2) Each record contains a unique set of SHAP values that are unique to it [26].

SHAP specifies the explanation for an instance x as:

$$g(z') = \varnothing 0 + \sum_{j=1}^{M} \varnothing_j z'_j. \tag{1}$$

The explanations of Eq. (1) are as follows: (1) g is the explanation model; (2) The coalition vector is represented by the symbol z' and is called a simplified feature; (3) $z' \in \{0, 1\}^M$, 1 indicates that the characteristics of the new data are identical to those of the original data (the instance x). In contrast, the value 0 indicates that the attributes in the new data are distinct from those in the original data (the instance x); (4) M is the maximum coalition size; (5) $\varnothing_j \in \mathbb{R}$ is the feature attribution for feature j, for instance, x. It is the Shapley value. If $\varnothing_j$ is a large positive number, it means feature j has an enormous positive impact on the prediction made by the model. SHAP assigns a weight to the sampled instances based on the coalition's weight in the Shapley value estimate process. Lundberg *et al.* proposed the SHAP kernel in Eq. (2).

$$\pi_x(z') = \frac{(M-1)}{(\frac{M}{|z'|})|z'|(M-|z'|)} \tag{2}$$

where M is the optimum coalition size and $|z'|$ is the total number of the entry of 1 in instance z'.

SHAP values have three major benefits over other techniques when compared to other methods. In the first place, SHAP has a strong theoretical basis in the field of game theory. Among the solutions available, symmetry, dummy, and additivity are the only three characteristics that may be satisfied by Shapley values [27]. SHAP may also meet these requirements since it obtains Shapley values from linear models. Second, SHAP establishes a connection between Local Interpretable Model-Agnostic Explanations (LIME) and Shapley values. It contributes to the consolidation of the area of interpretable machine learning. Finally, compared to simply computing the Shapley value, SHAP provides a faster calculation for machine learning models.

Fig. 3 illustrates the XAI captioning with SHAP architecture. In our experiment, we used 60 images from ImageNet to test the image captioning. Image Captioning based on meta-learning image understanding and visual concept. Next, the system generates the text, caption re-ranking, and final caption for each image. We use two ways to describe Image Captioning (Image to Text) as follows: (1) Azure Cognitive Services and (2) Open-Source Image Captioning Models.

2.3 Microsoft Azure Cognitive Service and Open-Source Image Captioning Model

Microsoft Cognitive Services is a comprehensive collection of Intelligence APIs that can be easily integrated into any application, according to Microsoft. Microsoft Cognitive Services, formerly known as Project Oxford, is built on the Azure Machine Learning platform (ML) [28]. Cognitive Services contains highly complicated, state-

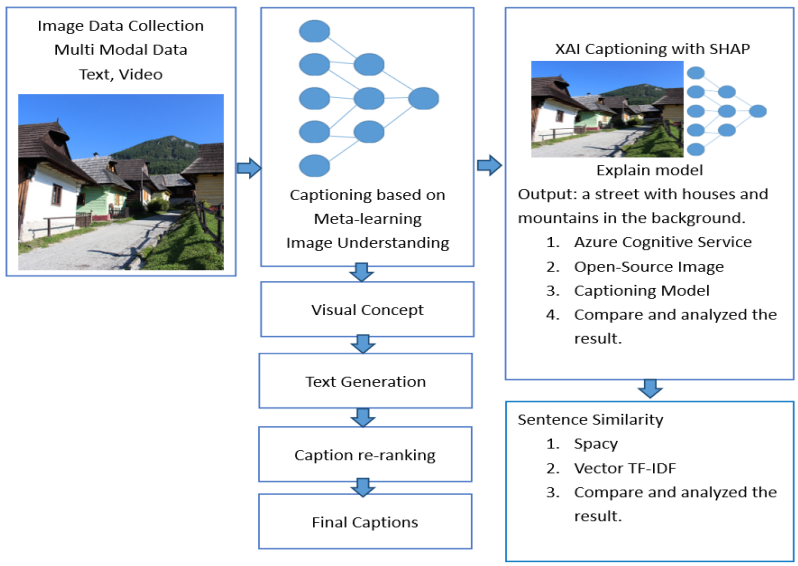


Fig. 3. XAI captioning with SHAP and sentence similarity architecture.

```
# place your Azure COGS CV subscription API Key and Endpoint below
API_KEY = "<your COGS API access key>"
ENDPOINT = "<endpoint specific to your subscription>"

ANALYZE_URL = ENDPOINT + "/vision/v3.1/analyze"
```

Fig. 4. API details.

of-the-art, Intelligence machine learning algorithms that are exposed as uniform and simple-to-use REST APIs and are available as SDKs for a limited number of different programming languages. Using REST APIs is simple and can be implemented in any type of application written in any language by modifying a few lines of code [29]. The Cognitive Services APIs are grouped into five categories as follows [30]; (1) Vision: Image analysis software extracts content and other valuable information from images and videos; (2) Speech: tools for enhancing voice recognition and establishing the speaker's identity; (3) Language: It is more important to understand phrases and meaning than simply words; (4) Knowledge: collects research from scholarly publications for the benefit of the user; (5) Search: machine learning is used for online searches. Furthermore, to use Azure Cognitive Services, obtain the API Key and Endpoint associated with our Azure Cognitive Services subscription. It is recommended to purchase a premium service instead of a free service to avoid rate caps on API calls and get a brief explanation. API details are shown in Fig. 4.

Azure Cognitive Services supports the following image file types: JPEG (JPG), PNG, GIF, BMP, and JFIF. Further, Cognitive Services have a maximum file size of 4MB and a minimum picture size of 50×50. Our experiments deformed large image files

to improve SHAP annotation performance and ran Azure Cognitive Services for image captions. If the image dimensions (pixel size, pixel size) are more than 500, the image is scaled to have a maximum pixel size of 500. The other dimensions are adjusted to maintain the original aspect ratio. The second way is to explain Image Captioning using the Open-Source Image Captioning Model. Our experiment used a pre-trained open-source model from R. Luo *et al.*, [31] to get image captions, and all pre-trained models are available. Moreover, our research experiment uses the model trained with ResNet101. Our works segment images along axes, for example, super pixels or partitions of halves, quarters, eights) to explain image captions. SHAP practices transformer language model Distil BERT [32] to adjust scoring within the given image and masked image captions. Assuming an external model is a better surrogate for the initial captioning model's language head. The better surrogate provides the most meaningful explanation for the image. By using the captioning model's language head, we could eliminate this assumption and remove the dependency. The greater the number of judgments required to generate annotations, the longer it will take SHAP to execute. However, an increase in the number of assessments refined explanations (300-500 evaluations often yielded detailed maps, but less or more often made sense).

2.4 Sentence Similarity Evaluations

Sentence similarity is a challenging research task with applications across many Natural Language Processing (NLP) tasks [33], such as summarizing documents, answering questions, and sentence generation. Phenomenal similarity and pragmatic similarity are two metrics to measure the distance between sentences. Syntactic similarity measures the similarity of the word structure of the phrase, while pragmatic similarity measures the similarity of context [34]. Spacy is a free open-source package for natural language processing (NLP) in Python that offers word vectors. It is built on the very latest state-of-the-art researches. It comes with a set of pre-trained statistical models and word vectors. It supports tokenization for more than 60 languages. Furthermore, Spacy uses cosine similarity by default. Word vectors or word embedding, multi-dimensional semantic representations of a word, are used to determine similarity. Cosine similarity is determined in Eq. (3).

$$similarity = \cos\theta = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

To determine how similar two words are semantical, a value between 0 and 1 is used. This is achieved by comparing word vectors in a vector space and determining how similar they are. spaCy, one of the fastest NLP libraries in use today, provides a direct method for this purpose. Hence, spaCy supports two methods to find word similarity: using context-sensitive tensors, and using word vectors. In our experiment, we implement *en_core_web_sm* English package. *en_core_web_sm* is a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities. This English pipeline is CPU optimized and the components include *tok2vec*, *tagger*, *parser*, *flashlight*, *ner*, *attribute_ruler*, *lemmatizer*. The well-known Term Frequency Inverse Document Frequency (TF-IDF transform) provides a good estimate of

the specified metric, and there are efficient implementations of it. The TF-IDF is a modification on the top of the TF as it downscales the weights for the terms that appear in several documents in the corpus and are therefore less informative compared to those that exist only in a lower percentage of the corpus and are more useful. TF is the term frequency, *i.e.* the frequency of the word *t* in document *d*, this is calculated in log space and shown in Eq. (4) [35]:

$$tf_{t,d} = \log_{10}(count(t, d) + 1), \tag{4}$$

$$idf_t = \log_{10}(\frac{N}{df_t}). \tag{5}$$

IDF is the inverse document frequency *N*/*df*; where *N* is the total number of documents in the collection, and df is the number of documents a term occurs in. This gives a higher weight to words that occur only in a few documents. Terms that are limited to a few documents are useful for discriminating those documents from the rest of the collection. The fewer the documents in which a term occurs, the higher this weight, this is also calculated in log space: TF-IDF is determined in Eq. (6).

$$\text{TF-IDF} = w_{t,d} = tf_{t,d} \times idf_t \tag{6}$$

To evaluate our sentence similarity method, we used conventional performance indicators, notably the F1 and accuracy scores, along with their related class support divisions. Precision and recall are defined in Eqs. (7) and (8). Moreover, accuracy and F1 are defined in Eqs. (9) and (10) [36].

$$Precision = \frac{TP}{TP+FP}, \tag{7}$$

$$Recall = \frac{TP}{TP+FN}, \tag{8}$$

$$Accuracy\ score = \frac{TP+TN}{TP+TN+FP+FN}, \tag{9}$$

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

Where True Positive (TP) is the number of reviews sorted properly into the appropriate sentiment classifications. Next, False Positive (FP) is the number of reviews assigned to an emotion class to which they do not belong. Hence, False Negative (FP) is the number of reviews labeled as not belonging to a sentiment category in which they really fit.

3. RESULTS AND DISCUSSION

3.1 SHAP Experiment Results

Table 1 explains our SHAP 3 image captioning result. Our experiment used 60 images and generated the image caption using (1) Azure Cognitive Service and (2) Open-Source Image Captioning Model. We compared the captions generated by the two meth-

ods and classified them into Yes and No classes. The classification process is done manually, one by one, by looking at the pictures and descriptions. If the image description is close to true, then it belongs to class Yes. If not means belongs to class No. In Table 1, Azure Cognitive Service generates an image caption "a bird perched on a branch," and it belongs to the Yes class. Next, the Open-Source Image Captioning Model produces the caption "a bird sitting on top of a tree branch" with class Yes. Although the two models produce slightly different descriptions, they have the same meaning in the Yes class.

Table 1. SHAP image captioning result.

Image	Image Size	(1) Azure Cognitive Services		(2) Open-Source Image Captioning Model	
		Image Caption	Class	Image Caption	Class
1	(500, 500, 3)	a bird perched on a branch	Yes	a bird sitting on top of a tree branch	Yes
2	(500, 500, 3)	a close up of an owl	Yes	a bird is standing on top of a tree	No
3	(500, 500, 3)	a turtle on the ground	No	a bird is sitting on top of a field	No

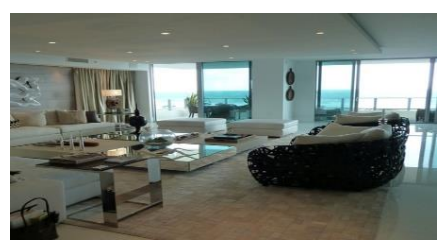
SHAP image captioning performance describes in Table 2. Furthermore, Azure Cognitive Services shows the highest accuracy of 87%. On the other hand, the Open-Source Image Captioning Model only reached 52% accuracy. Out of a total of 60 images, Azure Cognitive Services correctly generated 52 images captions.

Table 2. SHAP image captioning performance.

Model	Yes	No	Total	% Yes	% No	% Total
(1) Azure Cognitive Services	52	8	60	87	13	100
(2) Open-Source Image Captioning Model	31	29	60	52	48	100

(a)

(1) A snowy forest with trees.
(2) A tree is standing next to a tree with a tree.



(b)

(1) A living room with a large window.
(2) A living room with a couch and a table.

Fig. 5. Image captioning result class (1) Yes and (2) Yes.

Fig. 5 illustrates the results of image captioning with Yes and Yes classes. Although the image captions are slightly different but still have a suitable meaning and all models provide correct image captions. Fig. 5 (a) describes the image caption "a snowy forest with trees" and "a tree is standing next to a tree with a tree." Moreover, Fig. 5 (b) explains the image caption as (1) a living room with a large window and (2) a living room with a

(a)

(1) A yellow flower with a dark background.
(2) A vase with flowers is sitting on top of a table.

(b)

(1) A meerkat is looking up.
(2) A bird is sitting on top of a tree.

Fig. 6. Image captioning result class (1) Yes and (2) No.

couch and a table. Fig. 6 describes the results of image captioning with Yes and No classes. Azure Cognitive Services generate a correct image caption, and Open-Source Image Captioning Model generates the wrong image caption. As seen in Fig. 6 (a), the picture description is as follows: (1) a yellow flower with a dark background and (2) a vase with flowers sitting on top of a table. The image caption in Fig. 6 (b) is described as (1) a meerkat looking up and (2) a bird is sitting on a tree.

There are some explainer options in SHAP such as "*inpaint_ns*" and "*blur (kernel_xsize, kernel_xsize)*, and "*inpaint_telea*". In our experiment, image masker uses a blurring technique called "*blur (kernel_xsize, kernel_xsize)*". The recommended number of evaluations is 300-500, based on the SHAP tutorial, to get the explanations with sufficient granularity for the superpixels. More the number of evaluations, more the granularity but also increases run-time. A *flip argsort* sliced by four has been used to get SHAP values because we want to get the top 4 most probable classes for each image top 4 classes with decreasing probability.

Interpretation of SHAP output explanation describes in Fig. 7. In Fig. 7, the first image is classified as a seashore, with the next probable classes being a jigsaw puzzle, castle, and promontory. The second image is classified as a solar dish, followed by a radio telescope, fountain, and stage. The third image is classified as a bittern, with the next probable classes being chickadee, junco, and bustard. We can see the region of the bird appropriately highlighted in red super pixels.

Fig. 7. SHAP explanation for test images.

3.2 Sentence Similarity Evaluations Results

Our research experiment implements Cosine similarity and Term Frequency Inverse Document Frequency (TF-IDF transform) to evaluate the sentence similarity. We evaluate all text generation by SHAP one by one.

The range value similarity by spaCy is 0.34 to 0.94 and TF-IDF range value between 0.965 to 0.127. Information retrieval and topic analysis are two disciplines in which the TF-IDF is commonly used to evaluate the relevance of words in documents. TF-IDF is calculated using two indices: TF (term frequency) and IDF (interval distribution function) (reverse document frequency). The TF–IDF score is the sum of the TF and IDF scores. Moreover, spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. We process image captions for each model and calculate sentence similarity. Our experiment labelled "Yes" as "1" and "no" as "0". After calculating the similarity values with spaCy and TF-IDF, we labelled "values ≥ 0.60" as "1" and "values < 0.60" as "0". Table 3 describes sentence similarity process.

Table 3. Sentence similarity process.

Class Azure Cognitive Services (Sentence 1)	Open-Source Image Captioning Model (Sentence 2)	Similarity (Spacy)	Similarity (TF-IDF)	SpaCy	TF-IDF	Sentence 1	Sentence 2
a bird perched on a branch	a bird sitting on top of a tree branch	0.758	0.769	1	1	1	1
a close up of an owl	a bird is standing on top of a tree	0.549	0.595	0	0	1	0
a turtle on the ground	a bird is sitting on top of a field	0.596	0.599	0	0	0	0

Table 4. Statistic performance of similarity value comparison.

Items	Class Azure Cognitive Services				Open-Source Image Captioning Model			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
	Similarity Spacy				Similarity Spacy			
FALSE	1	0.33	0.5	24	0.67	0.55	0.6	29
TRUE	0.69	1	0.82	36	0.64	0.64	0.69	31
Accuracy			0.73	60			0.65	60
Macro avg	0.85	0.67	0.66	60	0.65	0.65	0.65	60
Weighted avg	0.82	0.73	0.69	60	0.65	0.65	0.65	60
	Similarity TF-IDF				Similarity TF-IDF			
Positive	0.17	0.12	0.14	8	1	0.21	0.34	29
Negative	0.87	0.9	0.89	28	0.57	1	0.73	31
Accuracy			0.8	60			0.62	60
	0.52	0.51	0.51	60	0.79	0.6	0.54	60
Weighted avg	0.78	0.8	0.79	60	0.78	0.62	0.54	60

Table 4 explains the statistical performance of similarity value comparison for each class. In Class Azure Cognitive Services, similarity according to spaCy achieves 73% accuracy, and similarity with TF-IDF shows 80%. Meanwhile, the similarity value of the Open-source Image Captioning Model with a spaCy of 65% and the similarity value of TF-IDF is 62%. The average similarity accuracy according to spaCy is 69% and with TF-IDF is 71%. Based on our experimental results, we can conclude that the similarity with TF-IDF is better than spaCy. Some advantages of TF-IDF are as follows; (1) The TF-IDF model contains information on the more important and the less important words.

Words with a higher score are more important, and those with a lower score are less important; (2) TF-IDF usually performs better in machine learning models and is easy to compute; (3) TF-IDF have some basic metric to extract the most descriptive terms in a document, and it can easily compute the similarity between 2 documents using it.

4. CONCLUSIONS

This work generates image captions using Azure Cognitive Service and Open-Source Image Captioning Model. We analyzed and discussed in detail the experiment result for each method. Based on the experiment result, we can conclude as follows: (1) We found Azure Cognitive Service gives the most meaningful explanations for images compared to Open-Source Image Captioning Model; (2) Our research work discovered that Azure Cognitive Service provided the best relevant explanations for images and achieved the maximum accuracy of 87%; (3) SHAP can explain the top 4 most probable classes for each image in our experiment; (4) Based on our experimental results, we can conclude that the similarity with TF-IDF is better than spaCy. In our future work, we will explore Azure Cognitive Services, especially for voice services API, to build voice recognition applications to facilitate visually impaired and visually impaired (BVI) with the XAI approach.

ACKNOWLEDGMENT

This paper is supported by the Ministry of Science and Technology, Taiwan. The Nos. are MOST-110-2927-I-324-50, MOST-110-2221-E-324-010, and MOST-109-2622-E-324-004, Taiwan. Additionally, this study was partially funded by the EU Horizon 2020 program RISE Project ULTRACEPT under Grant 778062.

REFERENCES

1. X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, and W. Dong, "Image caption generation with part of speech guidance," *Pattern Recognition Letters*, Vol. 119, 2019, pp. 229-237.
2. J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognition*, Vol. 98, 2020, doi: 10.1016/j.patcog.2019.107075.
3. K. Lalitha and S. Murugavalli, "A survey on image retrieval techniques," *Advanced Parallel Computing*, Vol. 37, 2020, pp. 396-400.
4. C. Dewi, R.-C. Chen, Y.-T. Liu, and S.-K. Tai, "Synthetic data generation using DCGAN for improved traffic sign recognition," *Neural Computing and Applications*, Vol. 33, 2021, pp. 1-15.
5. S. Khan, S. Nazir, and H. U. Khan, "Analysis of navigation assistants for blind and visually impaired people: A systematic review," *IEEE Access*, Vol. 9, 2021, pp. 26712-26734.
6. W. Hu *et al.*, "A comparative study in real-time scene sonification for visually impaired people," *Sensors*, Vol. 20, 2020, doi: 10.3390/s20113222.
7. C. Dewi, R. C. Chen, and H. Yu, "Weight analysis for various prohibitory sign de-

tection and recognition using deep learning," *Multimedia Tools and Applications*, Vol. 79, 2020, pp. 32897-32915.
8. C. Dewi, R.-C. Chen, and S.-K. Tai, "Evaluation of robust spatial pyramid pooling based on convolutional neural network for traffic sign recognition system," *Electronics*, Vol. 9, 2020, p. 889.
9. S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Transactions on Interactive Intelligent Systems*, Vol. 11, 2021, doi: 10.1145/3387166.
10. C. Dewi, R.-C. Chen, X. Jiang, and H. Yu, "Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks," *PeerJ Computer Science*, Vol. 8, 2022, https://doi.org/10.7717/peerj-cs.943.
11. C. S. Kumar, M. N. S. Choudary, V. B. Bommineni, G. Tarun, and T. Anjali, "Dimensionality reduction based on SHAP analysis: A simple and trustworthy approach," in *Proceedings of International Conference on Communication and Signal Processing*, 2020, pp. 558-560.
12. J. B. Lamy, K. Sedki, and R. Tsopra, "Explainable decision support through the learning and visualization of preferences from a formal ontology of antibiotic treatments," *Journal of Biomedical Science*, Vol. 104, 2020.
13. L. Zhang, K. Wu, B. Yang, H. Tang, and Z. Zhu, "Exploring virtual environments by visually impaired using a mixed reality cane without visual feedback," in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality Adjunct*, 2020, pp. 51-56,
14. C. Dewi, R.-C. Chen, X. Jiang, and H. Yu, "Deep convolutional neural network for enhancing traffic sign recognition developed on Yolo V4," *Multimedia Tools and Applications*, Vol. 81, 2022, pp. 37821-37845.
15. V. Lounnas *et al.*, "Visually impaired researchers get their hands on quantum chemistry: Application to a computational study on the isomerization of a sterol," *Journal of Computer-Aided Molecular Design*, Vol. 28, 2014, pp. 1057-1067.
16. X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (CBIR)," *Neurocomputing*, Vol. 452, 2021, pp. 675-689.
17. G. Kulkarni *et al.*, "Baby talk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, 2013, pp. 2891-2903.
18. J. Devlin, *et al.*, "Language models for image captioning: The quirks and what works," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Vol. 2, 2015, pp. 100-105.
19. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," *arXiv Preprint*, 2015, arXiv:1412.6632.
20. V. Atliha and D. Šešok, "Text augmentation using BERT for image captioning," *Applied Sciences*, Vol. 10, 2020, No. 5978.
21. O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, Vol. 115, 2015, doi: 10.1007/s11263-015-0816-y.
22. C. Schorr, P. Goodarzi, F. Chen, and T. Dahmen, "Neuroscope: An explainable AI toolbox for semantic segmentation and image classification of convolutional neural

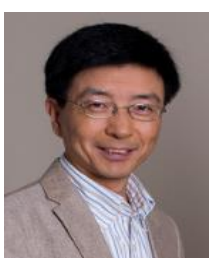
nets," *Applied Sciences*, Vol. 11, 2021, pp. 1-16.
23. S. H. Han and H. J. Choi, "Explainable image caption generator using attention and Bayesian inference," in *Proceedings of International Conference on Computational Science and Computational Intelligence*, 2018, pp. 478-481.
24. A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI," *Information Fusion*, Vol. 71, 2021, pp. 28-37.
25. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, Vol. 2017-Decem, 2017, pp. 4766-4775.
26. M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for Intrusion Detection Systems," *IEEE Access*, Vol. 8, 2020, pp. 73127-73141.
27. M. J. Ariza-Garzon, J. Arroyo, A. Caparrini, and M. J. Segovia-Vargas, "Explainability of a machine learning granting scoring model in peer-to-peer lending," *IEEE Access*, Vol. 8, 2020, pp. 64873-64890.
28. C. Dewi, R. C. Chen, H. Yu, and X. Jiang, "Robust detection method for improving small traffic sign recognition based on spatial pyramid pooling," *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, 2021, pp. 1-18.
29. S. Machiraju and R. Modi, "Azure cognitive services," in *Developing Bots with Microsoft Bots Framework*, Apress, 2018, pp. 233-260.
30. Microsoft Azure, "Microsoft Azure cognitive services," *Microsoft Docs*, 2021.
31. R. Luo, G. Shakhnarovich, S. Cohen, and B. Price, "Discriminability objective for training descriptive captions," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6964-6974.
32. A. Alshanqiti, A. Namoun, A. Alsughayyir, A. M. Mashraqi, A. R. Gilal, and S. S. Albouq, "Leveraging DistilBERT for summarizing Arabic text: An extractive dual-stage approach," *IEEE Access*, Vol. 9, 2021, pp. 135594-135607.
33. A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, 2021, pp. 4291-4308.
34. M. Farouk, "Measuring text similarity based on structure and word embedding," *Cognitive Systems Research*, Vol. 63, 2020, pp. 1-10.
35. H. M. Balaha and M. M. Saafan, "Automatic exam correction framework (AECF) for the MCQS, essays, and equations matching," *IEEE Access*, Vol. 9, 2021, pp. 32368-32389.
36. C. Dewi and R.-C. Chen, "Random forest and support vector machine on features selection for regression analysis," *International Journal of Innovative Computing, Information and Control*, Vol. 15, 2019, pp. 2027-2038.

Christine Dewi received a BS degree (S. Kom.) from the Informatics Engineering study program in 2010, and a Master of Computer Science (M.Cs.) from the Master of Information Systems study program in 2012, both from the Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia. In 2021 she finished her Ph.D. and she is now a Researcher at the College of Informatics, Chaoyang University of Technology,

Taiwan. She got Excellent Dissertation Awards from the Taiwanese Association for Consumer Electronics, Taiwan 2021. This award aims to honor a dissertation project funded by MOST Taiwan. Her current research interests include image processing, computer vision, object detection and recognition, artificial intelligence, and machine learning.

Rung-Ching Chen received a BS from the Department of Electrical Engineering in 1987, and an MS from the Institute of Computer Engineering in 1990, both from National Taiwan University of Science and Technology, Taipei, Taiwan. He received his Ph.D. from the Department of Applied Mathematics in computer science, National Chung Hsing University in 1998. Further, He is now a Distinguished Professor in the Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan. Also, His research concerns including network technology, pattern recognition, knowledge engineering, the Internet of Things, data analysis, and Artificial Intelligence.

Hui Yu is a Professor at the CCI Faculty of the University of Portsmouth. His Ph.D. work won the Best Ph.D. Thesis Prize, EPSRC DHPA Awards and Vice-Chancellor Travel Prize, *etc.* He previously held a research appointment with the University of Glasgow. Moreover, his research interests include vision, computer graphics, and application of machine learning to these areas, particularly in research areas such as image/video processing and analysis, 3D/4D sensing, reconstruction and geometric processing, human motion understanding, and effective analysis.



Xiaoyi Jiang received a bachelor's degree in Computer Science from Peking University, Beijing, China, and the Ph.D. and Venia Docendi (Habilitation) degrees in computer science from the University of Bern, Bern, Switzerland, in 1989 and 1997, respectively. He was an Associate Professor with the Technical University of Berlin, Berlin, Germany. Since 2002, he has been a Full Professor of Computer Science with the University of Münster, Münster, Germany, where he is currently the Dean with the Faculty of Mathematics and Computer Science. Dr. Jiang is a Fellow of International Association for Pattern Recognition. He is currently the Editor-in-Chief for the International Journal of Pattern Recognition and Artificial Intelligence. Besides, he also serves on the Advisory Board and Editorial Board of several journals, including IEEE Transactions on Medical Imaging and International Journal of Neural Systems.