International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

A Mobile Application Framework to Classify Philippine Currency Images to Audio Labels Using Deep Learning

Mary Grace Abellano Buban
College of Information Technology and Computer Science
University of the Cordilleras
Baguio City, Philippines
marygracebuban1117@gmail.com

Joyce Cadiz Malubay
College of Information Technology and Computer Science
University of the Cordilleras
Baguio City, Philippines
joycemalubay@gmail.com

Natividad Ballesteros Concepcion
College of Information Technology and Computer Science
University of the Cordilleras
Baguio City, Philippines
nbconcepcion@uc-bcf.edu.ph

Lilibeth Abellano Buban
Department of Education
School Division Office of Albay
Legazpi City, Philippines
lilibeth.buban@deped.gov.ph

Abstract— This research presents a mobile application framework designed to empower visually impaired individuals in Legazpi City by providing real-time audio feedback for currency identification. Leveraging deep learning techniques, the proposed framework employs a robust model trained on a comprehensive dataset of Philippine currency images. The deep learning model is capable of accurately classifying various denominations of bills and coins, enabling the development of an inclusive solution for the visually impaired community. The researcher employed a qualitative approach in this study, which included a focus group discussion. Respondents were chosen using purposive sampling. Among those who responded were masseuses, chiropractors, herbal street vendors, and students. Through an online meeting, the selected participants contributed to the focus group discussion. In addition, an in-depth informal interview was conducted to gather additional information for the development of an architectural framework. Based on the result of this study, it was discovered that by implementing this architectural framework, these groups would be able to more easily identify money, increasing efficiency and reducing errors in cash transactions. The use of audio labels is particularly helpful for visually impaired individuals, as it provides an accessible way for them to independently handle and identify money.

Keywords- Deep Learning, Image Captioning, Audio Mechanism, Android Application.

I. INTRODUCTION

In today's technologically advanced world, mobile-based applications have become a crucial tool in enhancing accessibility and improving the quality of life for various user groups, including visually impaired individuals. For the visually impaired community, identifying and recognizing currency notes is a significant challenge, as they heavily rely on tactile cues and assistance from others to determine the value of money. This limitation not only affects their financial independence but also restricts their overall mobility and participation in various economic activities. In Legazpi City, Albay, like many other urban areas, visually impaired individuals face these challenges on a daily basis. The absence of a reliable and accessible solution to accurately detect the currency denomination and read out its value in real-time significantly hampers their ability to carry out transactions independently and with confidence.

The utilization of Convolutional Neural Network (CNN) architectures for large-scale audio classification was explored. This underscored the efficacy of CNNs in processing audio signals for classification tasks, suggesting the potential adaptability of similar architectures for image data processing [1]. Furthermore, a unified framework, CNN-RNN, was proposed for multi-label image classification. This innovative approach seamlessly integrated CNNs and Recurrent Neural Networks (RNNs) to proficiently handle multi-label classification tasks. These findings advocate for the exploration of a hybrid CNN-RNN architecture for classifying currency images to audio labels, given the task's multi-class nature [2].

Insights into self-supervised visual feature learning with deep neural networks were provided, highlighting the significance of learning visual representations sans human annotations. This approach holds promise in developing unsupervised models for currency image classification [3]. Additionally, deep adversarial metric learning for cross-modal retrieval was introduced, focusing on learning a joint embedding space for different modalities. This technique holds relevance in aligning image and audio features for cross-modal classification, such as associating currency images with corresponding audio labels [4].

A study on audio-visual speech recognition employing deep learning demonstrated the efficacy of deep learning models in processing audio-visual inputs for speech recognition tasks. This insight could inform the development of a framework incorporating both visual (currency images) and audio (currency labels) inputs for classification [5]. Furthermore, design choices for deep audio embeddings were discussed, emphasizing the importance of feature representation and embedding design for audio classification tasks, which could inform the feature extraction process in the proposed framework [6]. Innovative image-to-audio translation using LSTM models was introduced, delving into the realm of cross-modal translation and enriching visual content interpretation [8].

Advancements in fintech were showcased through a study on currency note recognition, combining image processing with a minimum distance classifier to develop an automatic identification system [9]. Additionally, the potential of CNNs in detecting COVID-19 through chest X-ray analysis was discussed, highlighting deep learning's significance in medical imaging and disease diagnosis [10]. Environmental science benefited from deep learning research dedicated to mapping China's urban green spaces through satellite imagery analysis, aiding urban planning and environmental monitoring [11]. Novel image captioning and visual question answering methods employed both bottom-up and top-down attention mechanisms to refine tasks associated with image comprehension [12].

Challenges of limited labeled data for deep learning models were addressed through a semi-supervised learning strategy, combining CNNs with uncertainty filtering to enhance façade defect classification [13]. Further developments in remote sensing included a technique integrating superpixel pooling CNNs with transfer learning for hyperspectral image classification [15], while Unicoder-vl presented a universal encoder through cross-modal pre-training, pushing the envelope of vision and language task convergence [16]. In the automotive industry, deep learning showcased its revolutionary capabilities in tasks like autonomous driving, vehicle diagnostics, and predictive maintenance, marking a significant shift towards AI integration in the sector [17]. The adaptability of deep learning was evident in deep transfer learning for machine diagnosis, transferring knowledge from audio recognition to bearing fault detection and underscoring its versatility [18].

In bioacoustics, deep learning facilitated bird species classification through an augmented CNN, bolstering biodiversity monitoring [29]. A deep CNN-based strategy for animal sound classification tackled the complexities of acoustic environment analysis, aiding wildlife monitoring and ecological studies [30]. The intelligent diagnosis of brain tumors through deep CNNs and SVM algorithm integration in MRI analysis contributed significantly to medical image analysis, enhancing diagnostic and treatment planning accuracy [31]. Furthermore, attention-based models gained popularity in computer vision for generating natural language descriptions from images, setting new benchmarks for image captioning and visual question answering tasks [34].

The Attentional Generative Adversarial Network (AttnGAN) introduced attention-driven, multi-stage refinement for fine-grained text-to-image generation [35]. Additionally, the system for automatically generating natural language descriptions from images involved content planning and surface realization, enhancing the narrative of visual content [36]. Inspired by neuroscience, a recurrent CNN (RCNN) was proposed for object recognition by incorporating recurrent connections into each convolutional layer [37]. The Parti model was introduced to generate high-fidelity photorealistic images and support content-rich synthesis involving complex compositions and world knowledge [38]. Furthermore, to push the boundaries of vision-and-language pretraining data, the Conceptual 12M (CC12M) dataset with 12 million image-text pairs was introduced, specifically meant for vision-and-language pre-training [39]. The exploration of bi-directional mapping between images and their sentence-based descriptions dynamically builds a visual representation of the scene as a caption is generated or read [40].

City Social Welfare and Development Office Legazpi under the Department of Social Welfare and Development (DSWD), reported that the total number of households is 46, 445 and the registered visual disability as of September, 2023. There are 369 Males and 359 Females, for a total of 728. The majority of self-employed individuals with disabilities who earn an income in Legaspi City, Albay, have mobility impairments, compared to nearly half who are hearing-impaired. However, the group of people who depend the most on transfer income is the visually impaired, closely followed by the hearing-impaired. In addition, both the mobility- and hearing-impaired got the highest proportion of members who rely a lot on transfer income. The visually impaired, on the other hand, rely heavily on wages as well as self-employment, like masseurs. Therefore, this study may help most of the visually-impaired earners in identifying currency value. The challenges faced by the visually impaired in currency identification have led to research and development efforts aimed at addressing this issue. Various studies have explored the use of technology, particularly mobile applications, to provide assistance.

Therefore, this study presents the overall framework model to classify Philippine currency images to audio labels using deep learning using a learning-based convolutional neural network model running on a mobile application. The researcher also discusses the measurement requirements and classification process that can be used in designing a suitable application. The researcher believes that adapting the framework to the field of assistive technology for the visually impaired this study endeavors to create a robust and user-friendly mobile application that can potentially be extended to benefit visually impaired communities in other regions as well. By promoting financial independence and empowering the visually impaired with a tool that ensures seamless currency recognition and value comprehension, we envision a more inclusive and equitable

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

society, where every individual can participate actively in economic activities and lead a self-reliant life.

II. METHODS AND MATERIALS

A. Dataset Description

The researcher downloaded the Philippine Money datasets from the Roboflow website, as shown in Table 1, and captured some of them, resulting in 625 images. The researcher then created five captions for each image, resulting in 3,125 unique caption pairs, as shown in Table 2. Finally, datasets were loaded into COLAB in preparation for training, using the image processing technique. The goal of this study is to identify currency notes based on their denominations. The tflite model is important aspects of these tools provide a classified model for the project. In addition, the researcher used the CNN classification algorithm during the classification process. The process is followed by the capture of currency and then recognition output in the form of voice The model created serves not only as a recognizer but also as a calculator, as it automatically adds all of the currency denominations it recognizes, eliminating the need to manually enter them. gTTS makes it simple to add speech functionality to Python scripts, allowing you to generate audio files from any text voice.

TABLE I. TOTAL DATASETS FROM DOWNLOADED AND CAPTURED IMAGES

Denomination	Datasets		
	Downloaded	***Captured***	***Total***
1 cent	0	2	2
25 cent	14	10	24
1 peso	14	20	34
5 pesos	14	18	32
10 pesos	20	28	48
20 pesos	32	69	101
50 pesos	42	71	113
100 pesos	23	62	85
200 pesos	34	67	101
500 pesos	32	20	52
1000 pesos	23	10	33

Table 1 displays specific datasets, the number of images downloaded and captured, and the total number of images obtained.

TABLE II. TABLE TYPE STYLES

Denomination	Datasets		
	Total Images	***Caption per Image***	***Image with unique captions***
1 cent	2	5	10
25 cent	24	5	120
1 peso	34	5	170
5 pesos	32	5	160
10 pesos	48	5	240
20 pesos	101	5	505
50 pesos	113	5	565
100 pesos	85	5	425
200 pesos	101	5	505
500 pesos	52	5	260
1000 pesos	33	5	165

Table 2 displays the total image and is multiplied by 5 with a unique caption, resulting in a total of 3,125.

B. Components of Model

- gTTS (Google Text-to-Speech), a Python library and CLI tool to interface with Google Translate's text-to-speech API. Writes spoken mp3 data to a file, a file-like object (bytestring) for further audio manipulation, or stdout.
- Google Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education.
- TensorFlow is defined as an open-source platform and framework for machine learning, which includes libraries and tools based on Python and Java — designed with the objective of training machine learning and deep learning models on data.
- Keras is a neural network Application Programming Interface (API) for Python that is tightly integrated with TensorFlow, which is used to build machine learning models. Keras' models offer a simple, user-friendly way to define a neural network, which will then be built for you by TensorFlow.Units

C. Hardware and Software Specifications

The research used different libraries including OpenCV (cv2 version 4.8.0) for image processing, TensorFlow (version 2.15.0) for model simulation and validation, and Matplotlib (version 3.5.2) for visualizing results. It experimented with four optimization techniques (SGDM, RMSProp, Nadam, and Adam) on Google Colab's deep learning server using TensorFlow and Keras. The hardware setup included an Nvidia GeForce RTX 3050 Laptop GPU (version 512.78) and an AMD Ryzen 7 4800H CPU 8GB with 2.90 GHz. Mobile application testing was performed on a Realme GT Master Edition Android smartphone equipped with a Qualcomm Snapdragon 778G with Qualcomm Adreno 642L GPU.

III. RESULTS AND DISCUSSION

This chapter contains a detailed presentation and discussion of data gathered from participants in focus group discussions and informal interviews. As well as the measurement requirements and classification process.

Classifying or identifying money is important for everyone, it is especially important for those who are blind or visually impaired and can only do so by touching it. The main issue that these people are dealing with is determining the true amount of single money given to them. The majority of these people work as masseuses, chiropractors, herbal street vendors, and students. Therefore, it matters most for them that the money given is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

exactly right. In addition, some people take advantage of their disability and give them money that differs from the amount received.

Type of Works	Responses	Affected (%)
masseuses	uncertain about the salary received	45%
chiropractors and herbal street vendor	others take advantage of not paying the exact amount.	30%
students	unsure about the change given to them by a driver and eatery	25%

Figure 1. Existing Issues/Problem Faced by a Visually Impaired Person.

According to the data in the figure above, 45% of visually impaired masseuses are unsure of the salary received. Some people take advantage of not paying the exact amount for the purchased products, according to chiropractors and herbal street vendors with 30%. Some visually impaired students with 25% are also unsure of the change given by a driver and an eatery.

Person Involved	Response
Department of Social Welfare and Development (DSWD)	only offer limited assistance to workers and students with visual impairment only offer non-technological assistance.
masseuses, chiropractors, and herbal street vendor	although assistance is helpful, they believe they are being treated unfairly in terms of monetary compensation.
students	daily transportation difficulties without assistance from other passengers

Figure 2. Existing Issues/Problems Encountered by the existing Persons Involved.

Base on Figure 2, some of the problems encountered by CSWDO-Legazpi includes only limited assistance offered for workers and students and can only offers non-technological assistance. Also, some masseuses, chiropractors, and herbal street vendor stated that although assistance is helpful, they still feel they are being treated unfairly in terms of monetary compensation. The students with same situation also mention that the daily transportation is difficult for them without assistance from other passengers.

A. Images and Text Preprocessing

After loading the images and creating caption text for each value in the dataset, the researcher applied some preprocessing to both the images and the text. For images, the researcher used a feature vector for image preprocessing which is a list of numbers used to characterize the contents of an image, (see Figure 3). The researcher also uses the text preprocessing to converts the text to a lower string, removed punctuation and extra spaces, and inserts start and end special tokens to indicate the beginning and end of a sentence. Word cloud is also used by the researcher to provide an excellent option for analyzing text data through visualization in the form of tags or words, where the importance of a word is explained by its frequency. (see Fig. 3).

Figure 3. Word Cloud

The researcher use image preprocessing techniques such as noise reduction, contrast enhancement, image resizing, color correction, segmentation, and feature extraction to improve image quality and thus make them more suitable for analysis and further processing. Similar to the research, this study focuses on preserving the intrinsic structure of both images and text, resulting in more meaningful and context-aware representations. The use of deep learning in this context improves the ability to capture complex relationships between images and text [41].

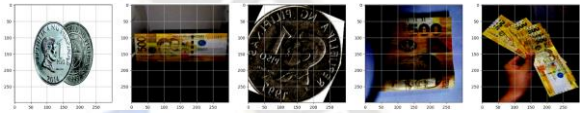

Figure 4. Image Preprocessing

Figure 4 shows the preprocessed images use in this study. In addition, one notable study investigated high-resolution image synthesis using latent diffusion models and demonstrated the potential of attention mechanisms in producing high-quality, realistic images. Similarly, other research indicates that semantic attention can effectively capture the salient regions of an image, resulting in more accurate image descriptions. Furthermore, this adaptive attention mechanism demonstrated promising results in creating more contextually relevant image captions. Another study showed that attention mechanisms can generate high-quality, realistic images. On the other hand, the study proposed a method for effectively inpainting missing regions in images using contextual attention. Other studies on image segmentation have demonstrated that scale-aware attention mechanisms can improve the performance of image segmentation algorithms by adaptively focusing on regions of varying scales [42], [43], [44], [45], [46], [47].

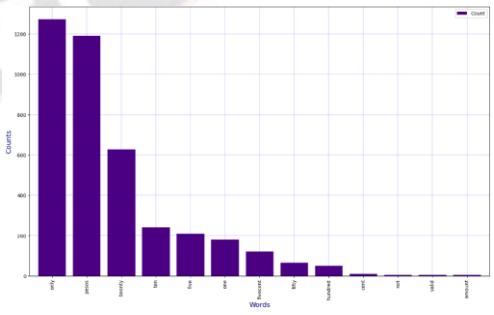

Figure 5. Maximum Frequency Words

The researcher uses graphical representations of word frequency shows in the figure above to highlight words that appear more frequently in a source text.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

IV. Proposed Architectural Framework

A control flow diagram is a tool that can aid visually impaired people in grasping intricate procedures. It shows the points where control commences and concludes, and the potential paths for divergence. The challenge is greater because the intended audience is blind and unaware of the conditions of their surroundings, including other objects, lighting, contrast, and whether the object of interest is within the camera's view. The system should be capable of handling a variety of images that the user might capture. If the program cannot identify the currency, the user must continue scanning until the currency is detected and announced.

Figure 6. Flowchart of the proposed emobile application framework

Figure 6 depicts the framework's flow, which begins with Scan Money and proceeds to pre-processing. If the datasets are not matched or found, an error message in audio format is sent, and the money must be scanned again. If a match is found in the datasets, the denomination is recognized, and the output in audio format is automatically displayed.

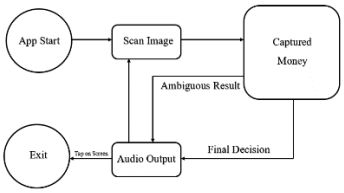

Figure 7. Control flow diagram of the proposed mobile application framework

Figure 7 depicts how the app will begin by scanning the image and then capturing money. If the result is unclear, the audio output will scan the image again, and if the captured money has the final decision, the audio output will be displayed.

Conclusion

This system is designed to assist various groups of people, including the Department of Social Welfare and Development (DSWD), masseuses, chiropractors, herbal street vendors, and students. These are groups that might frequently handle cash transactions and could benefit from an easier method of identifying banknotes. The framework was presented because it can assist with the following:

1. Masseuses and chiropractors often work in low-light environments where it can be difficult to distinguish between different banknotes.
2. Herbal street vendors and students might handle a high volume of cash transactions and need to quickly and accurately identify banknotes.
3. The DSWD could use this technology to assist visually impaired individuals in their care.

By implementing this architectural framework, these groups would be able to more easily identify money, increasing efficiency and reducing errors in cash transactions. The use of audio labels is particularly helpful for visually impaired individuals, as it provides an accessible way for them to independently handle and identify money.

Acknowledgment

The authors would like to express their gratitude to Department of Social Welfare and Development (DSWD) for their assistance with data collection; Prof. Natividad B. Concepcion, our research adviser, for her guidance; Joyce C. Malubay, whose insightful comments significantly enhanced the paper's quality; Jeffrey Ingosan. for fostering and assessing the abilities of the authors; and to our family, who provides us with spiritual support. We also express our gratitude to three anonymous reviewers whose suggestions helped to improve the quality of the paper. The writers gratefully acknowledge financial support from SIKAP Philippines' CHEDRO5.

References

[1] Hershey, Shawn., Chaudhuri, Sourish., Ellis, D.., Gemmeke, J.., Jansen, A.., Moore, R. C.., Plakal, Manoj., Platt, D.., Saurous, R.., Seybold, Bryan., Slaney, M.., Weiss, Ron J.., & Wilson, K.. (2016). CNN architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 131-135 . http://doi.org/10.1109/ICASSP.2017.7952132

[2] Wang, Jiang., Yang, Yi., Mao, Junhua., Huang, Zhiheng., Huang, Chang., & Xu, W.. (2016). CNN-RNN: A Unified Framework for Multi-label Image Classification. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2285-2294 . http://doi.org/10.1109/CVPR.2016.251

[3] Jing, Longlong., & Tian, Yingli. (2019). Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence , 43 , 4037-4058 . http://doi.org/10.1109/TPAMI.2020.2992393

[4] Xu, Xing., He, Li., Lu, Huimin., Gao, Lianli., & Ji, Yanli. (2019). Deep adversarial metric learning for cross-modal retrieval. World Wide Web , 22 , 657-672 . http://doi.org/10.1007/s11280-018-0541-x

[5] Noda, K.., Yamaguchi, Yuki., Nakadai, K.., Okuno, HIroshi G.., & Ogata, T.. (2014). Audio-visual speech recognition using deep learning. Applied Intelligence , 42 , 722 - 737 . http://doi.org/10.1007/s10489-014-0629-7

[6] Cramer, J.., Wu, Ho-Hsiang., Salamon, J.., & Bello, J.. (2019). Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 3852-3856 . http://doi.org/10.1109/ICASSP.2019.8682475

[7] Kumar, G., & Bhatia, P. (2014). A Detailed Review of Feature Extraction in Image Processing Systems. 2014 Fourth International Conference on Advanced Computing &

Communication Technologies, 5-12. http://doi.org/10.1109/ACCT.2014.74

[8] Zhang, H., Wang, Z., Huang, S., & Cai, Y. (2020). Image-to-audio Translation with Attention-based LSTM. Multimedia Tools and Applications, 79(35), 25675-25691.

[9] Sawant, K., & More, C. (2016). Currency Recognition Using Image Processing and Minimum Distance Classifier Technique. International Journal of Advanced Engineering Research and Science, 3(9), 236826.

[10] Bassi, P. R. A. S., & Attux, R. (2020). A deep convolutional neural network for COVID-19 detection using chest X-rays. Research on Biomedical Engineering, 38, 139-148. http://doi.org/10.1007/s42600-021-00132-9

[11] Shi, Q., Liu, M., Marinoni, A., & Liu, X. (2023). UGS-1m: fine-grained urban green space mapping of 31 major cities in China based on the deep learning framework. Earth System Science Data. http://doi.org/10.5194/essd-15-555-2023

[12] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2017). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077-6086. http://doi.org/10.1109/CVPR.2018.00636

[13] Guo, J., Wang, Q., & Li, Y. (2020). Semi‐supervised learning based on convolutional neural network and uncertainty filter for façade defects classification. Computer‐Aided Civil and Infrastructure Engineering, 36, 302-317. http://doi.org/10.1111/mice.12632

[14] Li, W., Wang, Z., Wang, Y., Wu, J., Wang, J., Jia, Y., & Gui, G. (2020). Classification of High-Spatial-Resolution Remote Sensing Scenes Method Using Transfer Learning and Deep Convolutional Neural Network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1986-1995. http://doi.org/10.1109/JSTARS.2020.2988477

[15] Xie, F., Gao, Q., Jin, C., & Zhao, F. (2021). Hyperspectral Image Classification Based on Superpixel Pooling Convolutional Neural Network with Transfer Learning. Remote Sens., 13, 930. http://doi.org/10.3390/rs13050930

[16] Li, Y., Huang, C., & Yang, L. (2020). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[17] Luckow, A., Cook, M., Ashcraft, N., Weill, E., Djerekarov, E., & Vorster, B. (2016). Deep learning in the automotive industry: Applications and tools. 2016 IEEE International Conference on Big Data (Big Data), 3759-3768. http://doi.org/10.1109/BigData.2016.7841045

[18] Brusa, E., Delprete, C., & Maggio, L. D. Di. (2021). Deep Transfer Learning for Machine Diagnosis: From Sound and Music Recognition to Bearing Fault Detection. Applied Sciences. http://doi.org/10.3390/app112411663

[19] Cui, M., Wu, G., Dang, J., Chen, Z., & Zhou, M. (2022). Deep learning-based condition assessment for bridge elastomeric bearings. Journal of Civil Structural Health Monitoring, 1-17. http://doi.org/10.1007/s13349-021-00540-6

[20] Nayyar, A., Pramanik, P. K., & Mohana, R. (2020). Introduction to the Special Issue on Evolving IoT and Cyber-Physical Systems: Advancements, Applications, and Solutions. Scalable Comput. Pract. Exp., 21, 347-348. http://doi.org/10.12694/scpe.v21i3.1568

[21] Janbi, N., & Almuaythir, N. (2023). BowlingDL: A Deep Learning-Based Bowling Players Pose Estimation and Classification. 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), 1-6. http://doi.org/10.1109/ICAISC56366.2023.10085434

[22] Xu, K. (2015, February 10). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv.org. https://arxiv.org/abs/1502.03044

[23] Kang, Y., Zhou, Y., Gao, M., Sun, Y., & Lyu, M. R. (2016, October 1). Experience Report: Detecting Poor-Responsive UI in Android Applications. https://doi.org/10.1109/issre.2016.16

[24] Wang, J., Huang, R., Guo, S., Li, L., Zhu, M., Yang, S., & Jiao, L. (2021). NAS-Guided Lightweight Multiscale Attention Fusion Network for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 59, 8754-8767. http://doi.org/10.1109/TGRS.2021.3049377

[25] Li, H., Cen, Y., Liu, Y., Chen, X., & Yu, Z. (2021). Different Input Resolutions and Arbitrary Output Resolution: A Meta Learning-Based Deep Framework for Infrared and Visible Image Fusion. IEEE Transactions on Image Processing, 30, 4070-4083. http://doi.org/10.1109/TIP.2021.3069339

[26] Wang, J., Zhang, J., Bao, W., Zhu, X., Cao, B., & Yu, P. S. (2018). Not Just Privacy: Improving Performance of Private Deep Learning in Mobile Cloud. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. http://doi.org/10.1145/3219819.3220106

[27] Yao, S., Zhao, Y., Hu, S., & Abdelzaher, T. (2018). QualityDeepSense: Quality-Aware Deep Learning Framework for Internet of Things Applications with Sensor-Temporal Attention. Proceedings of the 2nd International Workshop on Embedded and Mobile Deep Learning. http://doi.org/10.1145/3212725.3212729

[28] Hemelings, R., Elen, B., Barbosa-Breda, J., Lemmens, S., Meire, M., Pourjavan, S., ... Stalmans, I. (2020). Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. Acta Ophthalmologica, 98. http://doi.org/10.1111/aos.14193

[29] Jasim, H. A., Ahmed, S., Ibrahim, A., & Duru, A. (2022). Classify Bird Species Audio by Augment Convolutional Neural Network. 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 1-6. http://doi.org/10.1109/HORA55278.2022.9799968

[30] Oikarinen, T. P., Srinivasan, K., Meisner, O., Hyman, J., Parmar, S., Desimone, R., ... Feng, G. (2018). Deep Convolutional Network for Animal Sound Classification and Source Attribution using Dual Audio Recordings. bioRxiv. http://doi.org/10.1101/437004

[31] Wu, W., Li, D., Du, J., Gao, X., Gu, W., Zhao, F., ... Yan, H. (2020). An Intelligent Diagnosis Method of Brain MRI Tumor Segmentation Using Deep Convolutional Neural Network and SVM Algorithm. Computational and Mathematical Methods in Medicine, 2020. http://doi.org/10.1155/2020/6789306

[32] Lee, Hyungtae., & Kwon, H.. (2016). Going Deeper With Contextual CNN for Hyperspectral Image Classification. IEEE Transactions on Image Processing , 26 , 4843-4855 . http://doi.org/10.1109/TIP.2017.2725580

[33] Wei, Xiu-Shen., Luo, Jian-Hao., Wu, Jianxin., & Zhou, Zhi-Hua. (2016). Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. IEEE Transactions on Image Processing , 26 , 2868-2881 . http://doi.org/10.1109/TIP.2017.2688133

[34] Anderson, Peter., He, Xiaodong., Buehler, Chris., Teney, Damien., Johnson, Mark., Gould, Stephen., & Zhang, Lei. (2017). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 6077-6086 . http://doi.org/10.1109/CVPR.2018.00636

[35] Chen, Xinlei., & Zitnick, C. L.. (2015). Mind's eye: A recurrent visual representation for image caption generation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2422-2431 . http://doi.org/10.1109/CVPR.2015.7298856

[36] Xu, Tao., Zhang, Pengchuan., Huang, Qiuyuan., Zhang, Han., Gan, Zhe., Huang, Xiaolei., & He, Xiaodong. (2017). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 1316-1324 . http://doi.org/10.1109/CVPR.2018.00143

[37] Kulkarni, Girish., Premraj, Visruth., Dhar, Sagnik., Li, Siming., Choi, Yejin., Berg, A.., & Berg, Tamara L.. (2013).

BabyTalk: Understanding and Generating Simple Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence , 35 , 2891-2903 . http://doi.org/10.1109/TPAMI.2012.162

[38] Liang, Ming., & Hu, Xiaolin. (2015). Recurrent convolutional neural network for object recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 3367-3375 . http://doi.org/10.1109/CVPR.2015.7298958

[39] Yao, Ting., Pan, Yingwei., Li, Yehao., & Mei, Tao. (2018). Exploring Visual Relationship for Image Captioning. , 711-727 . http://doi.org/10.1007/978-3-030-01264-9_42

[40] Yu, Jiahui., Xu, Yuanzhong., Koh, Jing Yu., Luong, Thang., Baid, Gunjan., Wang, Zirui., Vasudevan, Vijay., Ku, Alexander., Yang, Yinfei., Ayan, Burcu Karagol., Hutchinson, B.., Han, Wei., Parekh, Zarana., Li, Xin., Zhang, Han., Baldridge, Jason., & Wu, Yonghui. (2022). Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. Trans. Mach. Learn. Res. , 2022 . http://doi.org/10.48550/arXiv.2206.10789

[41] Wang, Liwei., Li, Yin., & Lazebnik, S.. (2015). Learning Deep Structure-Preserving Image-Text Embeddings. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 5005-5013 . http://doi.org/10.1109/CVPR.2016.541

[42] Anderson, Peter., He, Xiaodong., Buehler, Chris., Teney, Damien., Johnson, Mark., Gould, Stephen., & Zhang, Lei. (2017). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 6077-6086 . http://doi.org/10.1109/CVPR.2018.00636

[43] You, Quanzeng., Jin, Hailin., Wang, Zhaowen., Fang, Chen., & Luo, Jiebo. (2016). Image Captioning with Semantic Attention. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 4651-4659 . http://doi.org/10.1109/CVPR.2016.503

[44] Lu, Jiasen., Xiong, Caiming., Parikh, Devi., & Socher, R.. (2016). Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 3242-3250 . http://doi.org/10.1109/CVPR.2017.345Wang

[45] Rombach, Robin., Blattmann, A.., Lorenz, Dominik., Esser, Patrick., & Ommer, B.. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 10674-10685 . http://doi.org/10.1109/CVPR52688.2022.01042

[46] Yu, Jiahui., Lin, Zhe L.., Yang, Jimei., Shen, Xiaohui., Lu, Xin., & Huang, Thomas S.. (2018). Generative Image Inpainting with Contextual Attention. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 5505-5514 . http://doi.org/10.1109/CVPR.2018.00577

[47] Chen, Liang-Chieh., Yang, Yi., Wang, Jiang., Xu, Wei., & Yuille, A.. (2015). Attention to Scale: Scale-Aware Semantic Image Segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 3640-3649 . http://doi.org/10.1109/CVPR.2016.396