Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages

Batyr Arystanbekov[1], Askat Kuzdeuov[1], Shakhizat Nurgaliyev[1], and Huseyin Atakan Varol[1]

***Abstract*— Visually impaired and blind people often face a range of socioeconomic problems that can make it difficult for them to live independently and participate fully in society. Advances in machine learning pave new venues to implement assistive devices for the visually impaired and blind. In this work, we combined image captioning and text-to-speech technologies to create an assistive device for the visually impaired and blind. Our system can provide the user with descriptive auditory feedback in the Kazakh language on a scene acquired in real-time by a head-mounted camera. The image captioning model for the Kazakh language provided satisfactory results in both quantitative metrics and subjective evaluation. Finally, experiments with a visually unimpaired blindfolded participant demonstrated the feasibility of our approach.**

I. Introduction

Image captioning is a task at the interface between computer vision and natural language processing that deals with creating a textual description of an image [1]. This involves processing an image, recognizing and understanding the objects, scenes, and activities depicted in it, and then using this information to generate a corresponding textual description. Image captioning has numerous applications in such areas as robotics, visual media retrieval, and accessibility.

Initial approaches that mostly relied on statistical [2] and graph-based [3] methods were followed by neural networks. These neural networks extracted rich visual features and produced text using encoder-decoder structures [4], [5], [6]. Convolutional neural networks were used to encode the visual information, while the decoding part leveraged recurrent neural networks. With the introduction of the attention mechanism [7], the performance of image captioning models improved drastically due to the efficient connection between the encoder and decoder parts. Later, attention-based models were further developed by the self-attention mechanism—that is, the application of attention inside the encoding and decoding parts [8]. Recently, an image captioning architecture with a new expansion mechanism has been proposed to utilize different sequence lengths in the encoder and decoder [9].

In 2017, the World Health Organization (WHO) estimated that 253 million people were visually impaired, of whom 36 million were blind [10]. Visually impaired and blind people suffer from severe socioeconomic disadvantages, such as unemployment, limited educational opportunities, social isolation, discrimination, and inaccessible technology. Presumably, these problems could be alleviated by assistive technologies.

Assistive technology refers to adaptive, and rehabilitative devices used by people with disabilities to perform tasks that they may have difficulty with or cannot do without someone else's support. Image captioning combined with text-to-speech technology can serve as an aid for the visually impaired and blind. For instance, verbal description of a scene acquired in real time can help visually impaired and blind people to determine their location, navigate their surroundings, access visual information, feel safer, and increase their situational awareness [11], [12]. Indeed, there are some recent works looking at the use of image captioning to help the visually impaired and blind [13], [14]. These studies focus on the English language, because benchmark image captioning datasets with human-generated captions (e.g., Microsoft COCO Captions dataset [15]) are generally created in this high-resource language.

There is already some work on image captioning for other languages (e.g., Italian [16] and Arabic [17]) using the Microsoft COCO Captions dataset with neural machine-translated versions of the captions. However, these are not integrated with text-to-speech technology to serve as assistive technology for the blind. In this work, we present the first assistive system for visually impaired and blind people that combines image captioning with text-to-speech technology for images acquired in real time for Kazakh—a low-resource language (see Fig. 1). Our work can serve as a recipe for the development of similar systems for other low-resource languages: less studied, resource scarce, less privileged [18]. To support further research in this area, the codes, dataset, and models used in our study are available for download from our GitHub repository[1].

The rest of the paper is organized as follows: In Section II, we describe the preparation process of the dataset, the architecture of the image captioning model, and the details of model training and real-world deployment. The results of image captioning in Kazakh for the COCO dataset and the real-world demonstration of the assistive system with a visually unimpaired blindfolded user are presented and discussed in Section III. Finally, in Section IV, the main conclusions of this work are drawn and areas for future research are suggested.

[1]B. Arystanbekov, A. Kuzdeuov, S. Nurgaliyev, and H.A. Varol are with the Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Astana, Kazakhstan. E-mail: {batyr.arystanbekov, askat.kuzdeuov, shakhizat.nurgaliyev, ahvarol}@nu.edu.kz.

[1]https://github.com/IS2AI/kaz-image-captioning

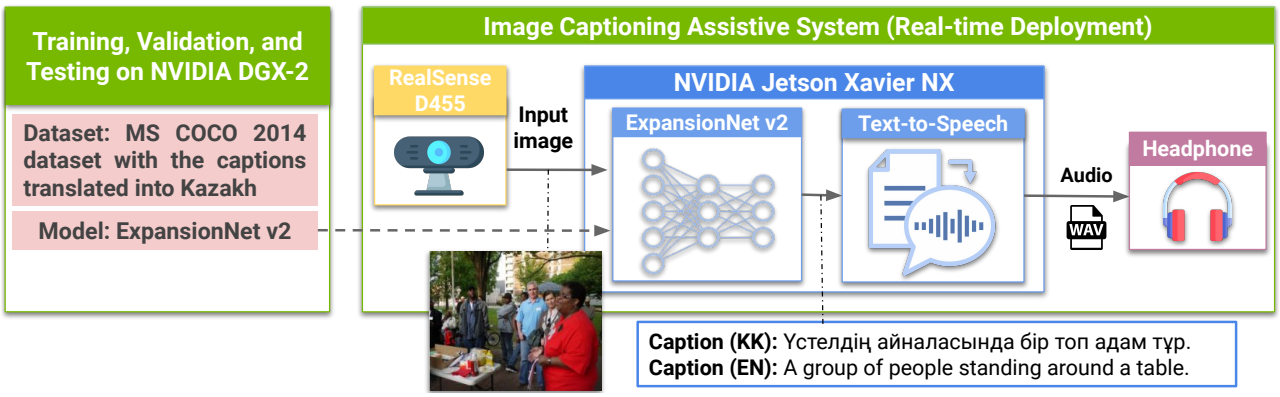

Fig. 1. Hardware and software architecture of our image captioning assistive system for model training and real-time deployment.

II. Methodology

A. Dataset Preprocessing

We used the popular Microsoft COCO 2014 (COCO) benchmark dataset [19] to train the ExpansionNet v2 image captioning model [20]. The dataset consisted of 123,287 images, with each image having five human-annotated captions, resulting in a total of over 600,000 image-text pairs. We split the dataset into training (113,287 images), validation (5,000 images), and test (5,000 images) sets, using the "Karpathy" splitting strategy [21] for offline evaluation. To generate captions in Kazakh, we translated the original English captions using the freely available Google Translate service. To evaluate the quality of the translated captions, 2,500 captions were translated from English into Kazakh by human translators. We then calculated the BLEU-4 [22] score for the machine-translated captions. We obtained a metric value of 0.49, which generally corresponds to a high-quality machine translation [23]. In addition, we preprocessed the captions by lowering casing, filtering out punctuation, and removing words with fewer than five occurrences, resulting in a vocabulary of 18,363 words.

B. Image Captioning Model

Because of its state-of-the-art results on the COCO dataset, we chose ExpansionNet v2 [20] as our image captioning model. To train the model for Kazakh captions, we followed the model architecture defined in the original work. The pretrained Swin Transformer [24] was used as a backbone network to generate visual features from the input images. The result was passed to the encoder, consisting of $N_{enc} = 3$ static expansion layers, while the decoder consisted of $N_{dec} = 3$ dynamic expansion layers. Finally, the output of the model was fed into the classification layer.

C. Model Training

We first preprocessed the input images for model training. The images were resized into $3 \times 384 \times 384$ tensors. Then, the RGB values were converted into a range of $[0, 1]$ and normalized by subtracting the mean $(0.485, 0.456, 0.406)$ and dividing by the standard deviation $(0.229, 0.224, 0.225)$.

The training procedure consisted of two phases. The first phase was the pretraining of the model, using the cross-entropy loss function. In the second phase, self-critical optimization was conducted. The two phases were further divided into two smaller steps: the initial training step, during which we froze the weights of the backbone (eight epochs in pretraining and nine epochs in reinforcement), and the fine-tuning step with end-to-end training (two epochs in pretraining and one in reinforcement). In each of the four training steps, the batch size was set to 48. For other hyperparameters, we refer the interested reader to the original paper [20]. The model was trained on four V100 graphics processing units (GPUs) in an Nvidia DGX-2 server.

D. Model Deployment

The architecture of our image captioning assistive system is shown in Fig. 1. The system consisted of a camera (Intel RealSense D455), a single-board deep learning computer (Nvidia Jetson Xavier NX), a push button, and headphones. The camera was connected to the single-board computer via universal serial bus (USB). The push button and headphones were connected to the general-purpose input/output (GPIO) pins and audio port of the single-board computer, respectively. The camera was attached to the user's forehead with adjustable straps. The user wore the headphones and carried the single-board computer (and a power bank) in a backpack.

The image captioning model, ExpansionNet v2, was deployed on the Nvidia Jetson Xavier NX (Nvidia Volta with 384 Nvidia CUDA and 48 Tensor cores, 6-core NVIDIA Carmel ARM v8.2 64-bit CPU). The camera was triggered by pressing the push button to capture an RGB image with a resolution of 640×480 pixels. Then, the captured image was resized to 384×384 and passed to the ExpansionNet v2 model to generate a caption. Next, the generated caption text was converted into audio, using a text-to-speech model. In this work, we used a rapid, offline TTS system - Piper [25]. The system provides models trained for the Kazakh language using the KazakhTTS dataset [26]. The synthesized audio was subsequently transmitted to the user via headphones.

III. Results and Discussion

A. Offline Evaluation on the COCO Dataset

To quantitatively evaluate our image captioning model, we employed metrics commonly used in machine translation, such as BLEU-4 [22], METEOR [27], and CIDEr [28].

Kazakh: Мотоцикл мінген адам қара жолда келе жатыр.
English: A man riding a motorcycle is riding on a dirt road.

Kazakh: Диван мен үстелі бар қонақ бөлмесі.
English: Living room with sofa and table.

Kazakh: Далада жайылып жүрген бір отар қой.
English: A flock of sheep grazing in the field.

Kazakh: Ұялы телефонмен сөйлесіп тұрған бір топ әйелдер.
English: A group of women talking on mobile phones.

Fig. 2. Sample images from the test set of the COCO dataset with the corresponding predicted captions in Kazakh and their English translations.

The results for the English and Kazakh models on the offline "Karpathy" test set are given in Table I. The original ExpansionNet v2 model [20] showed 41.0 BLEU-4, 30.2 METEOR, and 139.6 CIDEr scores for the original English captions. Indeed, it is possible to translate the predicted English captions into Kazakh instead of training the model with Kazakh captions. In this case, the translated captions yielded 20.0 BLEU-4, 20.8 METEOR, and 71.8 CIDEr scores for the translated Kazakh captions. In comparison, the ExpansionNet v2 model trained on Kazakh captions obtained higher scores for the translated Kazakh captions (21.9 BLEU-4, 21.8 METEOR, and 81.3 CIDEr).

TABLE I
MODEL EVALUATION ON THE COCO "KARPATHY" TEST SPLIT

Model	BLEU-4	METEOR	CIDEr
ExpansionNet v2 (en)	41.0	30.2	139.6
ExpansionNet v2 (en-kk)	20.0	20.8	71.8
ExpansionNet v2 (kk)	21.9	21.8	81.3

Sample images from the test set with the predicted captions in Kazakh and their English translations are shown in Fig. 2. The model provided accurate captions for the images from the different domains in Fig. 2a-c. However, the caption for Fig. 2d is slightly inaccurate, as the group of women is not talking on mobile phones. On the other hand, this caption can be considered partially correct, because the image does depict a group of women, one of whom is holding a mobile phone. In general, the model provides the overall context correctly, although in some cases it fails on certain details.

B. Real-World Experiments

We also conducted real-world experiments to test the efficacy of our assistive device. A visually unimpaired blindfolded participant was taken to various locations on a university campus. The participant used our assistive device to obtain information about the scene, as shown in Fig. 3a. In these experiments, we tried to consider cases that a blind person may face in their daily life. For instance, the image captioning system was able to correctly convey information about a staircase despite some redundancies in the statement (see Fig. 3b). We also obtained sufficiently accurate and useful captions for the scenes shown in Figs. 3c-e. The subject commented positively on the usefulness of the system and suggested that it would be desirable to obtain information about the text inside the signs. Presumably, our system can be integrated with an optical character recognition model to read aloud the text inside the detected signs.

The computation time of various modules within the system was also assessed to examine latency during real-world operations, with the outcomes presented in Table II. ExpansionNetV2's original code [20] stores the model utilizing a .pth file extension, resulting in a 2.7 GB model size and an average inference time of 1.34 s on the GPU of the NVIDIA Jetson. By optimizing the original model (.pth) and converting it to a TensorRT file format (.engine), the model size was reduced to 986 MB and the average inference time decreased to 0.49 s (standard deviation: 0.02 s) on the GPU. The average inference time of the Piper TTS model on a CPU of the NVIDIA Jetson was 0.73 s. This makes the computation time for a user request around 1.2 s, demonstrating real-time operation with low latency.

IV. CONCLUSION

In this paper, we have presented an assistive device for the visually impaired and blind that uses image captioning technology to provide descriptions of scenes in real time. Our method can serve as a recipe for such systems for low-resource languages for which there is no dedicated image caption database. The experiments with a visually unimpaired subject have proven the feasibility of our approach. In the future, we will work on deploying our system on smartphones, so that it can serve a wider community. In addition, we will explore how contextual information into our image captioning model can be embedded, so that relevant captions are generated for the visually impaired and blind.

TABLE II
INFERENCE TIME (SEC) OF EXPANSIONNET V2 (PYTORCH, TENSORRT) @384X384 PIXELS, AND PIPER TTS ON AN NVIDIA JETSON

# image	PyTorch, 2.7 GB	TensorRT, 986 MB	TTS, 18 MB
1	2.56	0.53	0.93
2	1.14	0.48	0.75
3	1.16	0.47	0.71
4	1.12	0.49	0.79
5	1.17	0.46	0.76
6	1.21	0.48	0.81
7	1.35	0.5	0.61
8	1.50	0.5	0.64
9	1.12	0.46	0.61
10	1.10	0.5	0.73
Mean	**1.34**	**0.49**	**0.73**
Std. dev.	**0.42**	**0.02**	**0.09**

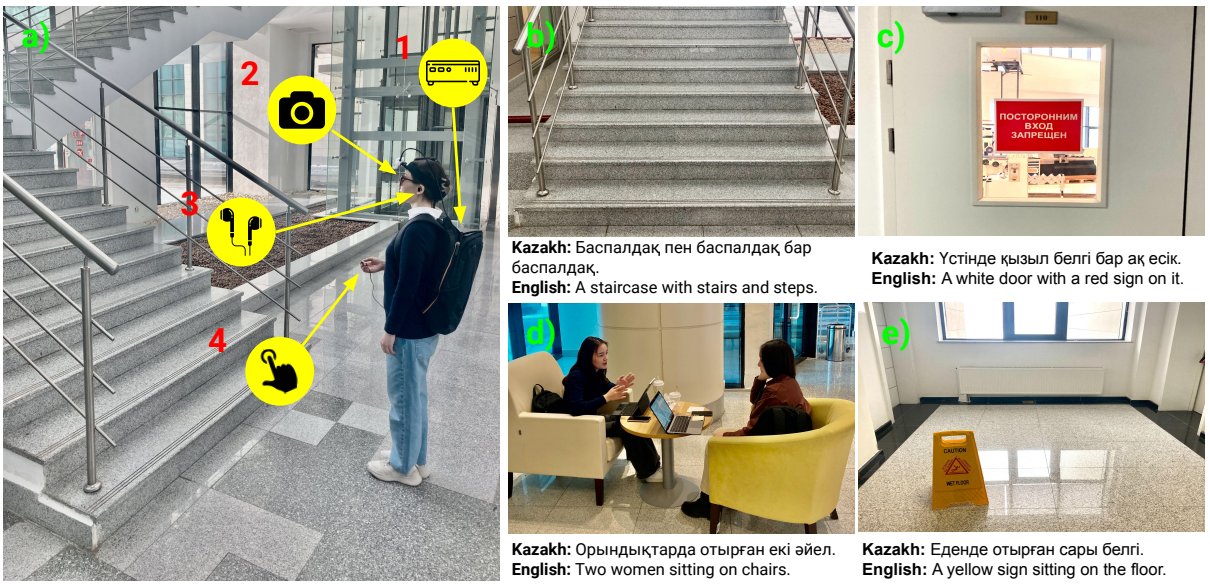

Fig. 3. a) A real-world experiment with a human subject wearing the image captioning assistive device: 1) Nvidia Jetson NX, 2) Intel RealSense D455, 3) headphones, and 4) push button. b-e) Four real-world images with the predicted captions in Kazakh and their English translations.

REFERENCES

[1] M. Stefanini, M. Cornia, L. Baraldi *et al.*, "From show to tell: A survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539–559, 2023.

[2] M. Mitchell, J. Dodge, A. Goyal *et al.*, "Midge: Generating image descriptions from computer vision detections," in *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, Apr. 2012, pp. 747–756.

[3] G. Kulkarni, V. Premraj, S. Dhar *et al.*, "Baby talk: Understanding and generating simple image descriptions," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1601–1608.

[4] P. Anderson, X. He, C. Buehler *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086.

[5] O. Vinyals, A. Toshev, S. Bengio *et al.*, "Show and tell: A neural image caption generator," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, 2017.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

[8] M. Cornia, M. Stefanini, L. Baraldi *et al.*, "Meshed-memory transformer for image captioning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[9] J. C. Hu, R. Cavicchioli, and A. Capotondi, "Exploring the sequence length bottleneck in the transformer for image captioning," 2022. [Online]. Available: https://arxiv.org/abs/2207.03327

[10] R. R. A. Bourne, S. R. Flaxman, T. Braithwaite *et al.*, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review," *The Lancet Global Health*, vol. 5, no. 9, pp. e888–e897, 2017.

[11] E. Brady, M. R. Morris, Y. Zhong *et al.*, "Visual challenges in the everyday lives of blind people," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, p. 2117–2126.

[12] A. Bhowmick and S. Hazarika, "An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends," *Journal on Multimodal User Interfaces*, vol. 11, pp. 1–24, 2017.

[13] J. Ganesan, A. T. Azar, S. Alsenan *et al.*, "Deep learning reader for visually impaired," *Electronics*, vol. 11, no. 20, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/20/3335

[14] B. Makav and V. Kılıç, "Smartphone-based image captioning for visually and hearing impaired," in *Proc. of the International Conference on Electrical and Electronics Engineering (ELECO)*, 2019, pp. 950–953.

[15] X. Chen, H. Fang, T.-Y. Lin *et al.*, "Microsoft COCO Captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

[16] C. Masotti, D. Croce, and R. Basili, "Deep learning for automatic image captioning in poor training conditions," in *Proc. of the CEUR Workshop*, 2017.

[17] H. A. Al-muzaini, T. N. Al-yahya, and H. Benhidour, "Automatic arabic image captioning using RNN-LSTM-based language model and CNN," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018.

[18] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *CoRR*, vol. abs/2006.07264, 2020.

[19] T. Lin, M. Maire, S. J. Belongie *et al.*, "Microsoft COCO: Common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[20] J. C. Hu, R. Cavicchioli, and A. Capotondi, "ExpansionNet v2: Block static expansion in fast end to end training for image captioning," 2022. [Online]. Available: https://arxiv.org/abs/2208.06551

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *CoRR*, vol. abs/1412.2306, 2014. [Online]. Available: http://arxiv.org/abs/1412.2306

[22] K. Papineni, S. Roukos, T. Ward *et al.*, "BLEU: A method for automatic evaluation of machine translation," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[23] A. Lavie, "Evaluating the output of machine translation systems," in *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts*, Xiamen, China, Sep. 19 2011.

[24] Z. Liu, Y. Lin, Y. Cao *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[25] M. Hansen, "Piper: A fast, local neural text to speech system," 2023. [Online]. Available: https://github.com/rhasspy/piper

[26] S. Mussakhojayeva, A. Janaliyeva, A. Mirzakhmetov *et al.*, "KazakhTTS: An open-source Kazakh text-to-speech synthesis dataset," in *Proc. of the Interspeech*, 2021, pp. 2786–2790.

[27] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, Jun. 2005, pp. 65–72.

[28] R. Vedantam, C. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.