УДК 004

ANALYSIS AND DEVELOPMENT OF A SYSTEM FOR OBJECT RECOGNITION AND DESCRIPTION IN IMAGES FOR MOBILE DEVICES

R. Z. Urazalin

Satbayev University Specialty, Almaty, Republic of Kazakhstan.

Introduction. Modern computer vision technologies are actively applied in various fields - from industry to medical diagnostics. One of the most significant areas is assisting visually impaired people who need to receive up-to-date information about their surroundings. Thanks to the rapid development of mobile computing resources, it has become feasible to offload complex computational tasks to smartphones, allowing users to obtain scene descriptions in near real-time [1]. The goal is to provide a comprehensive, non-visual understanding of the immediate environment, replicating, to some extent, the visual perception of sighted individuals. This capability is paramount for independent navigation and interaction.

1. Object Recognition with Deep Neural Networks. The core of the object recognition system relies on the application of Deep Convolutional Neural Networks (DCNNs). Such sophisticated models are capable of simultaneously determining the coordinates of objects in an image (localization) and assigning their correct class label (classification). Unlike traditional algorithms that heavily relied on laborious, hand-crafted features (like SIFT or HOG), modern models leverage extensive training on large-scale datasets (e.g., MS COCO, ImageNet), which significantly improves recognition accuracy and robustness across varied lighting and viewing conditions.

Specifically, popular detection architectures often include Single Shot Detectors (SSDs) or You Only Look Once (YOLO) variants, known for their speed, or two-stage detectors like Faster R-CNN, which offer superior precision. The detector processes an image streamed from a mobile device camera and

identifies a set of objects with bounding boxes that uniquely identify the position and type of objects in the scene. The extracted features - such as object class, confidence score, and relative spatial coordinates - are then formalized into a feature vector and transferred to the subsequent text generation module. The overall architectural flow, from image acquisition to text generation, is conceptualized in Figure 1.

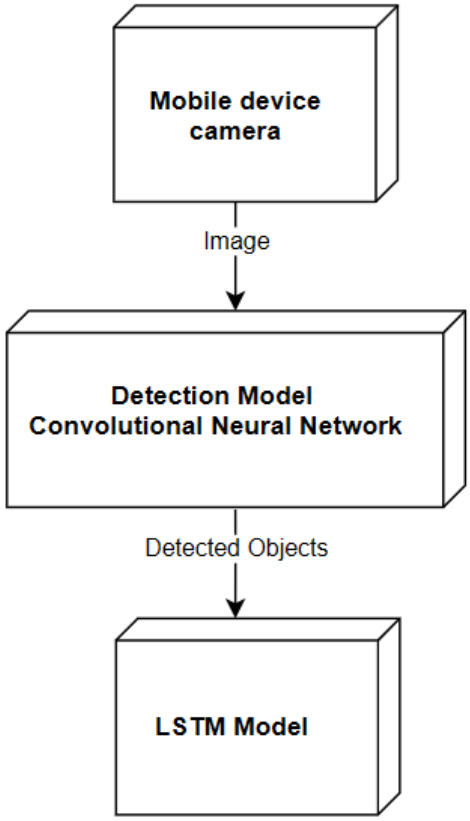

Figure 1 - Object Recognition and Description Scheme

2. Scene Description Generation via Recurrent Models. As depicted in Figure 1, text description generation, often termed Image Captioning, is a separate, complex task solved using natural language processing (NLP) methods. The most common and effective architectural approach is the Encoder-Decoder framework, where the DCNN acts as the Encoder (extracting visual features), and a Recurrent Neural Network (RNN) acts as the Decoder (generating text).

The Long Short-Term Memory (LSTM) architecture [2] is particularly favored for the decoder due to its inherent ability to handle sequential data processing effectively. LSTM units are specifically designed with internal 'gates' (input, forget, and output) to mitigate the vanishing gradient problem, enabling them to capture long-range dependencies crucial for forming grammatically correct and logically coherent sentences [3].

The LSTM model receives the formalized features obtained after object detection as input: object classes, their estimated relative positions, and the count of specific objects. The model then forms a coherent text where scene objects are described taking into account contextual relationships and spatial logic. For instance, instead of just listing objects, it can generate: "A cup is on the table, next to a book," providing crucial relational context.

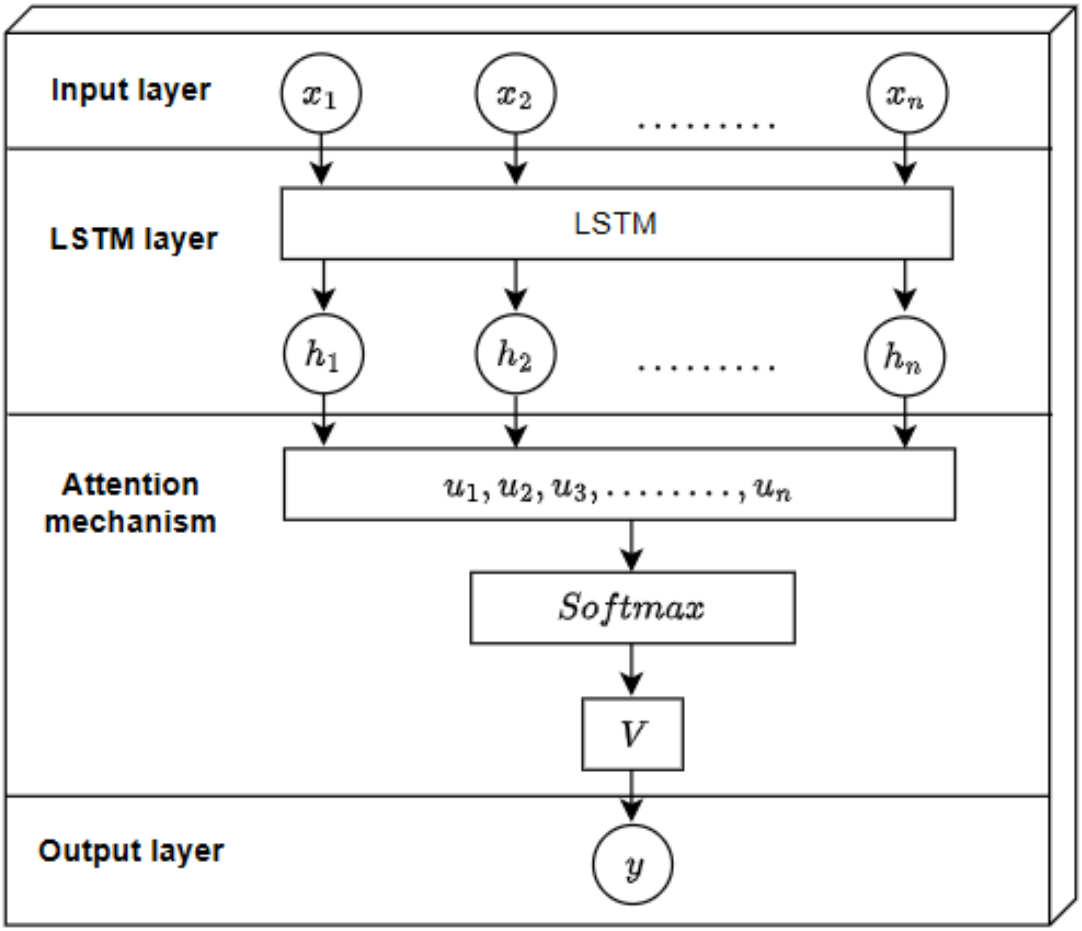

Figure 2 - Structure of the Attention Mechanism Layer

3. The Role of Attention Mechanisms. As illustrated in Figure 2, in modern systems, attention mechanisms are frequently applied. [4]. This sophisticated technique allows the recurrent decoder model to dynamically focus on the most significant regions of the input image (or corresponding object features) at each step of the word generation process. For example, when the model generates the word "cup," the attention mechanism will put a higher weight on the visual features of the cup object. Thanks to this mechanism, the scene description becomes more precise, informative, and contextually relevant, directly benefiting the end-user. For visually impaired users, the final description can be delivered either in text or immediately converted to an audio format using high-quality speech synthesis technologies; ensuring information is perceived without any need for visual contact.

4. Mobile Platform Advantages and System Deployment. The choice of the mobile platform offers several distinct advantages. Firstly, mobile devices (smartphones and tablets) are ubiquitous and virtually always at hand, ensuring the system is highly accessible for every day, spontaneous use. Secondly, modern smartphones are equipped with high-quality; high-resolution cameras that provide acceptable input image quality essential for accurate DCNN processing.

Crucially, the computational power of contemporary mobile platforms is continually increasing, allowing a significant portion of the processing to be carried out locally (on-device processing) or in a hybrid format [5]. On-device processing minimizes latency (response time) and maintains user privacy. However, in cases where more complex, state-of-the-art recognition and text generation models are required, part of the processing can be offloaded to a server via cloud computing. The mobile device simply captures the image and receives the description back, which effectively reduces the load on local resources and battery consumption.

5. Practical Applications and Future Directions. The use of these advanced technologies allows providing visually impaired users with critical real-time information about people, objects, road signs and vehicles, as well as the location

of objects in the immediate environment. The system proves invaluable in everyday scenarios, such as the automatic description of package contents, objects on a table surface, or the denomination of banknotes.

For successful practical deployment, response time (latency) is a critical parameter: the delay between capturing an image and forming the spoken description must be minimal to support dynamic interaction. Furthermore, the availability of training data covering diverse environments, lighting conditions, and object types is essential to improve the model's robustness and generalization capabilities.

Similar and highly effective functionality is already present in existing commercial solutions, such as Microsoft Seeing AI and Google Lookout, which provide users with voice descriptions of objects and text in images [6]. The continuous challenge remains the further improvement of description quality, which is inherently dependent on the sophistication of the underlying models and the breadth and composition of the training data.

The development of recognition systems and automatic description of objects in images represents a significant area for the further improvement of computer vision and AI accessibility technologies. Potential development directions include the integration of more powerful and context-aware transformer architectures (like the use of the Vision Transformer as an encoder), expansion of multi-modal and multi-lingual datasets, and adaptation of systems to various languages and cultural contexts [7]. Ultimately, these solutions are vital tools for enhancing the social inclusion and autonomy of people with vision impairment, effectively bridging the gap between the user and their physical environment.

References:

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

2. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
3. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164.
4. K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 2048-2057.
5. "Seeing AI: New Technology Research to Support the Blind and Visually Impaired Community," Microsoft Accessibility Blog, 2019. [Online]. Available: https://blogs.microsoft.com/accessibility/seeing-ai/
6. A.Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.
7. A. K. Muhammed Kunju, S. Baskar, S. Zafar et al., "A transformer based real-time photo captioning framework for visually impaired people with visual attention," Multimedia Tools and Applications, vol. 83, pp. 88859-88878, 2024.