



Fusion of High-Level Visual Attributes for Image Captioning

Murat Kılıcı¹, Özkan Çaylı¹, Volkan Kılıç¹

¹ İzmir Katip Çelebi University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, İzmir, Turkey, (ORCID: 0009-0000-3192-1601, 0000-0002-3389-3867, 0000-0002-3164-1981), 190403023@ogr.ikc.edu.tr, ozkan.cayli@ikcu.edu.tr, volkan.kilic@ikcu.edu.tr

(First received 18 August 2023 and in final form 10 September 2023)

(DOI: 10.5281/zenodo.10260172)

ATIF/REFERENCE: Kılıcı, M., Çaylı, Ö., & Kılıç, V., (2023). Fusion of High-Level Visual Attributes for Image Captioning. *European Journal of Science and Technology*, (52), 161-168.

Abstract

Image captioning aims to generate a natural language description that accurately conveys the content of an image. Recently, deep learning models have been used to extract visual attributes from images, enhancing the accuracy of captions. However, it is essential to assess these visual attributes to ensure optimal performance and avoid incorporating redundant or misleading information. In this study, we employ the visual attributes of semantic segmentation, object detection, instance segmentation, keypoint detection, and their fusion. Experimental evaluations on the commonly used datasets VizWiz and MSCOCO Captions demonstrate that the fusion of visual attributes improves the accuracy of caption generation. Furthermore, the image captioning model, which utilizes the fusion of visual attributes, has been embedded into our custom-designed Android application, named *NObstacle*, enabling captioning without the need for an internet connection.

Keywords: Visual Attributes, Image Captioning, Android Application.

Görüntü Altyazılama için Üst Düzey Görsel Özniteliklerin Birleştirilmesi

Öz

Görüntü altyazılama, bir görüntünün içeriğini doğru olarak ileten bir doğal dil açıklaması üretmeyi amaçlar. Son zamanlarda, altyazıların doğruluğunu arttırmak için görsel öznitelikleri çıkaran derin öğrenme modelleri kullanılmaktadır. Ancak, optimal performansın sağlanması, gereksiz ve yanıltıcı bilgilerin işlenmesinin önlenmesi açısından bu görsel özniteliklerin değerlendirilmesi önemlidir. Bu çalışmada, anlamsal bölümlenme, nesne algılama, örnek bölümlenme, anahtar nokta algılamanın ve bunların birleşiminin görsel öznitelikleri kullanıyoruz. VizWiz ve MSCOCO Captions gibi yaygın olarak kullanılan veri kümelerinde yapılan deneysel değerlendirmeler, görsel özniteliklerin birleşiminin altyazı üretiminin doğruluğunu artırdığını göstermektedir. Ayrıca, görsel özniteliklerin birleşimini kullanan görüntü altyazılama modeli, *NObstacle* adını verdiğimiz özel tasarlanmış Android uygulamamıza entegre edilerek internet bağlantısı gerektirmeden altyazı üretimini sağlamaktadır.

Anahtar Kelimeler: Görsel Öznitelikler, Görüntü Altyazılama, Android Uygulama.

1. Introduction

The task of image captioning aims to generate a meaningful and grammatically correct sentence to describe an image, which is achieved through the utilization of techniques from computer vision and natural language processing fields (Akosman et al., 2021; Fetiler et al., 2021; Sayraci et al., 2023). This task has found industrial and practical applications, such as visual question answering (Anderson et al., 2018; Keskin et al., 2021), image indexing (Baran et al., 2021; Chang, 1995), and virtual assistants (Doğan et al., 2022; Makav & Kılıç, 2019).

Recent studies mostly employ retrieval-based, template-based, and neural encoder-decoder-based frameworks in image captioning (Farhadi et al., 2010; Moral et al., 2022). The retrieval-based framework generates a candidate caption by presenting a similar image of a given dataset to the models. In the retrieval-based framework, a set of candidate captions is generated from the reference captions in the dataset, which are similar to the input image. From this candidate set, the caption that captures the most semantic information of the input image is chosen (Betül et al., 2022; Kılıç et al., 2014; Yang et al., 2020). In template-based methods, a caption is generated by matching the visual information of detected objects and actions with fixed templates (Kılıç, 2021; Yu et al., 2019).

The encoder-decoder framework was proposed to describe the contents of images because it captures representations of visual data as a latent vector (Farhadi et al., 2010; Mercan & Kılıç, 2020; Çaylı et al., 2021). Typically, the encoder utilizes a convolutional neural network (CNN), whereas the decoder employs a recurrent neural network (RNN) (Pu et al., 2020). The encoder represents the image as a latent vector that captures objects and semantic information. On the other hand, the decoder utilizes the image representation to generate a natural sentence (Çaylı et al., 2023; Jiang et al., 2018; Aydın et al., 2022).

Training deep learning models from scratch is a time-consuming and computationally costly process. Therefore, most studies generally prefer to utilize pre-trained deep-learning models in the encoder. This utilization offers significant contributions in terms of time and computational cost (Çaylı et al., 2022; Keskin et al., 2021; Makav & Kılıç, 2019). Visual attributes represent fundamental elements to describe different characteristics of an object, a scene, or visual content (Deselaers et al., 2004; Mercan et al., 2020). These attributes include features such as color, shape, size, location, or intensity of an object within a scene. In the field of computer vision, visual attributes play a significant role, serving as essential components for various tasks such as object detection (Amit et al., 2020), and instance segmentation (Ibarra et al., 2017; Kılıç et al., 2022; Liu et al., 2018). In semantic segmentation, each pixel in an image is classified based on its respective category, providing a comprehensive understanding of the image content (Wang et al., 2018). Instance segmentation is a computer vision task that goes beyond conventional semantic segmentation. While semantic segmentation assigns each pixel in an image to a specific category, instance segmentation advances this by identifying and outlining distinct instances of objects within each category (Liu et al., 2018). Object detection aims to identify and localize the objects within an image, based on their visual characteristics (Amit et al., 2020). Likewise, keypoint detection involves identifying essential points of interest in objects (e.g., facial features such as eyes, nose, and mouth) and determining their respective positions (Barroso-Laguna et al., 2019).

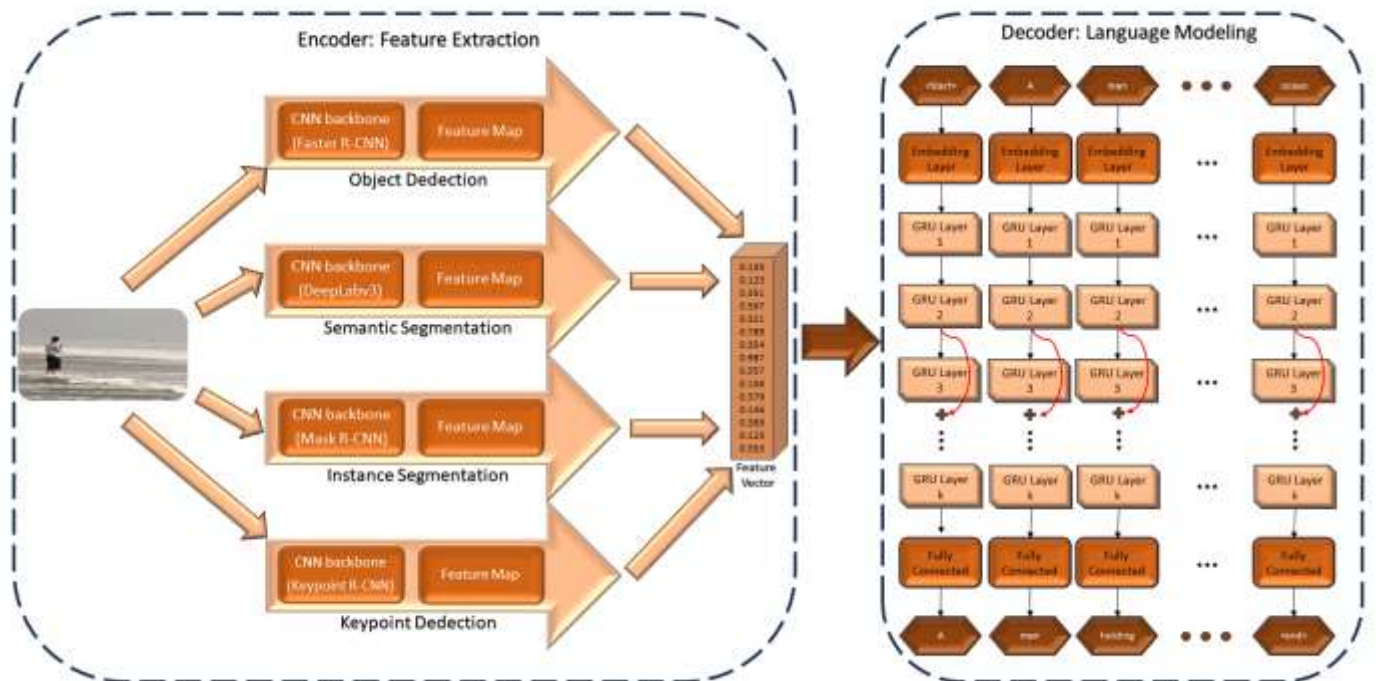


Figure 1 The Proposed Image Captioning Approach.

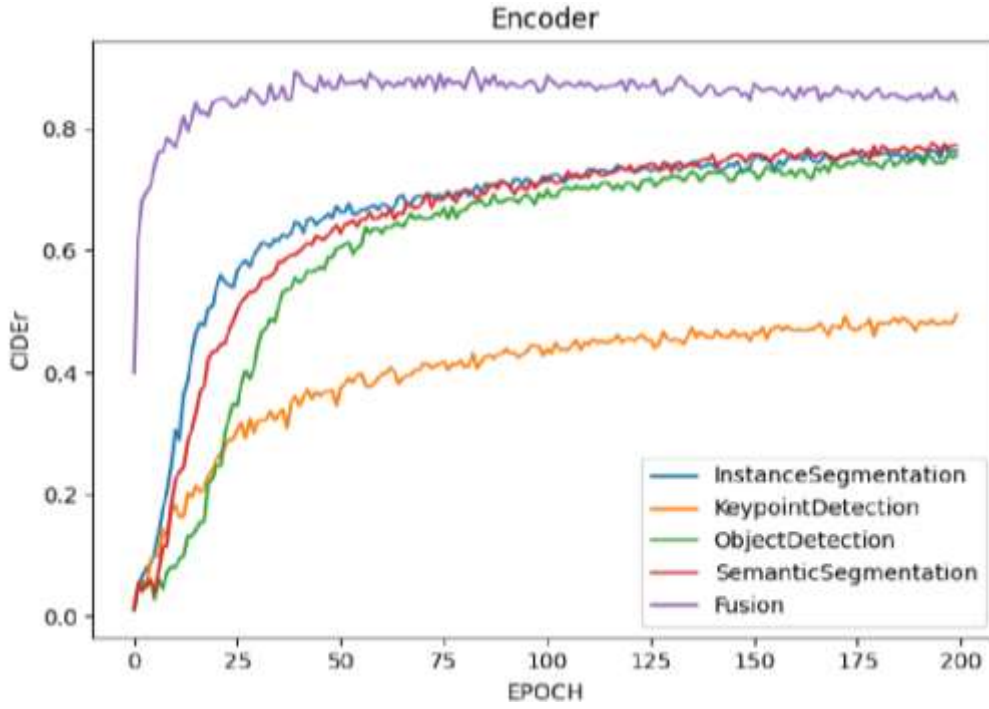


Figure 2 Evaluations of visual attributes with their fusion in terms of CIDEr on MSCOCO dataset.

In this study, we present a fusion of high-level visual attributes for more accurate and contextually relevant captioning by capturing more semantic information in an image. The approach utilizes pre-trained models based on ResNet in the encoder, incorporating Instance Segmentation, Semantic Segmentation, Object Detection, and Keypoint Detection techniques. The process of extracting attributes is performed using models such as DeepLabv3, Mask R-CNN, and a Feature Pyramid Network (FPN). The decoder utilizes the residual connected GRU model, providing gradient flow through residual connections between subsequent layers. We used the MSCOCO (Lin et al., 2014) and VizWiz (Gurari et al., 2020) datasets for experiments and evaluated the efficiency of the proposed approach with performance metrics CIDEr, SPICE, METEOR, ROUGE-L, and BLEU-n (n =1, 2, 3, 4).

The rest of the paper is organized as follows: Section 2 describes the proposed image captioning approach as shown in Figure 1, visual attributes extraction methods, and our custom-designed Android application *NObstacle*. Section 3 covers the dataset, evaluation metrics, and results. Finally, Section 4 concludes the study with closing remarks.

2. Methods

In this section, we first introduce the image captioning approach along with utilized visual attributes. Then, we present an Android application named *NObstacle* that is capable of generating captions offline.

2.1. The Proposed Approach

In the proposed approach, we extract semantic, object, instance, and keypoint attributes from a given image to generate a caption using an RNN-based decoder. These extracted attributes are utilized in single, pairwise, triplet, and quadruplet combinations to feed into the RNN-based decoder. The purpose of the decoder is to predict a caption word-by-word sequentially until an end-of-caption word is generated. Gated Recurrent Unit (GRU) (Chung et al., 2014) model is the type of RNN that employ a hidden state to retain and propagate information across sequential data points. GRU computes the hidden state h_s as shown in (1).

$$\begin{aligned}
 r_g &= \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}) \\
 u_g &= \sigma(W_{iu}x_t + b_{iu} + W_{hu}h_{t-1} + b_{hu}) \\
 n_g &= \tanh \tanh (W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{t-1} + b_{hn})) \\
 h_s &= (1 - z_t) \odot n_t + z_t \odot (h_{s-1})
 \end{aligned} \tag{1}$$

Table 1 PERFORMANCE METRIC RESULTS IS: Instance Segmentation, KD: Keypoint Detection, OD: Object Detection, SS: Semantic Segmentation (*: MSCOCO Captions Datasets, **: VizWiz Captions Datasets)

	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	SPICE	SCORE
Instance Segmentation (IS) *	0.769	0.243	0.344	0.486	0.668	0.491	0.222	0.149	0.413
Keypoint Detection (KD) *	0.495	0.188	0.272	0.403	0.595	0.438	0.181	0.104	0.317
Object Detection (OD) *	0.757	0.242	0.342	0.484	0.666	0.489	0.221	0.148	0.410
Semantic Segmentation (SS) *	0.775	0.252	0.352	0.494	0.676	0.495	0.223	0.150	0.417
IS+OD *	0.847	0.265	0.365	0.504	0.680	0.506	0.236	0.162	0.441
IS+KD *	0.811	0.254	0.355	0.497	0.676	0.500	0.229	0.158	0.429
IS+SS *	0.868	0.268	0.369	0.511	0.686	0.512	0.240	0.170	0.450
KD+OD *	0.839	0.266	0.367	0.508	0.684	0.506	0.234	0.161	0.439
KD+SS *	0.839	0.268	0.368	0.508	0.683	0.507	0.234	0.163	0.440
OD+SS *	0.888	0.279	0.383	0.524	0.700	0.518	0.242	0.169	0.458
IS+KD+OD *	0.849	0.266	0.368	0.509	0.685	0.508	0.237	0.163	0.443
IS+KD+SS *	0.869	0.271	0.371	0.512	0.685	0.512	0.241	0.170	0.450
IS+OD+SS *	0.890	0.278	0.379	0.519	0.693	0.516	0.242	0.171	0.457
KD+OD+SS *	0.889	0.275	0.379	0.520	0.695	0.516	0.243	0.170	0.457
Fusion *	0.898	0.280	0.382	0.523	0.695	0.515	0.241	0.172	0.459
Fusion **	0.298	0.157	0.245	0.379	0.577	0.394	0.159	0.089	0.256
(Chen et al., 2018)	0.600	0.121	0.191	0.308	0.505	-	-	-	-
(You et al., 2018)	0.665	0.136	0.207	0.322	0.511	0.390	0.170	-	-

The gates of a GRU, namely the reset, update, and new gate, can be denoted as follows: r_g , u_g , and n_g , respectively. In (1) subscript g means gate and i refers to the input. The Hadamard product is called \odot , while the sigmoid activation function is denoted as σ .

In the study, a multi-layered GRU-based model, which is specifically developed to work with sequential data, is utilized for caption generation. Utilizing the GRU-based model, the process of generating captions involves denoting the target sentence as $Y = y_1, y_2, \dots, y_N$, represents the sequence of words. Similarly, \hat{Y} represents sequential predictions of the network, respectively. To train the model, the cross-entropy (CE) loss is employed as the criterion, and the loss is computed as follows: $\text{loss} = \text{CE}(Y, \hat{Y})$.

2.2. Visual Attributes Extraction Methods

In this study, the DeepLabv3 architecture, which is trained for semantic segmentation tasks, has been employed to extract visual attributes. The objective of this task is to identify all corresponding pixels of objects within an image and allocate these pixels into discrete object categories (Guo et al., 2018). This task, commonly referred to as pixel-wise classification, serves the purpose of precisely segregating diverse objects presented within an image. The extraction of semantic segmentation attributes was accomplished by leveraging pre-trained weights from the MSCOCO dataset. Object detection is a computer vision task that aims to identify the positions and classes of objects within images (Amit et al., 2020). This involves segmenting the images into distinct areas, followed by examining each area to specify the location and type of the object (Tahir et al., 2021). Visual attributes from an object detection task are employed, which uses the Faster R-CNN architecture to extract attributes from images. Instance segmentation is a task to predict class labels and pixel-wise instance masks to accurately localize multiple instances within an image and leverages a Mask R-CNN (He et al., 2017; Liu et al., 2018). In this study, we employ ResNet50 CNN as the backbone layer to extract visual attributes (Xi et al., 2021). Keypoint detection task is typically used for tasks such as pose estimation in natural images, object pose estimation, and facial landmark localization. Human pose estimation involves predicting specific locations on the human body, like the elbow and wrist. There are primarily two approaches: directly determining the coordinates of these keypoints (Sun et al., 2018; Toshev & Szegedy, 2014) and producing heatmaps for keypoint detection. We utilized a visual attribute extraction process using pre-trained weights on the MSCOCO dataset (Lin et al., 2014).

2.3. Android Application: *NObstacle*

The proposed approach is integrated into the custom-designed Android application called *NObstacle*, which generates captions offline. Users can capture an image within the application using the camera or select from the gallery. Then, the proposed approach generates a corresponding caption and presents it to the user. Initially, the image captioning approach was fine-tuned using PyTorch (Paszke et al., 2019), an open-source framework developed for machine learning tasks on mobile devices. To minimize the weight of the model and thereby speed up the captioning process, we employed the dynamic quantization technique. In addition, the application supports various languages with speech command recognition. Screenshots of the *NObstacle* are given in Figure 3.

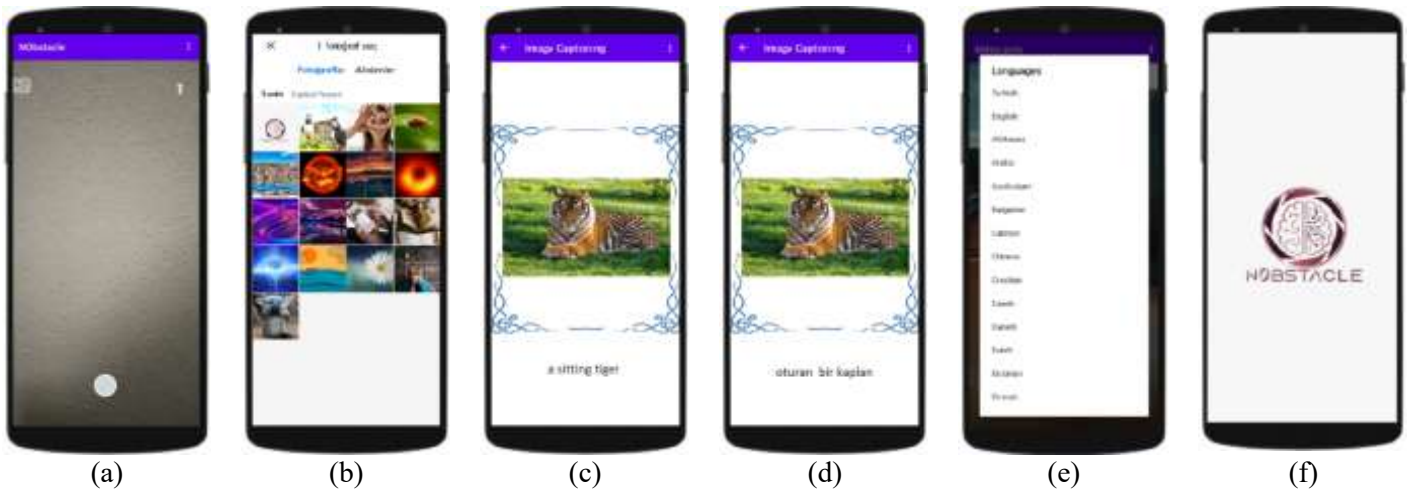


Figure 3 (a) shows the startup screen displays the application logo. (b) presents the home screen. (c) shows captions in English are visible on the main screen. (d) presents provided various language options. (e) shows an overview of the gallery. (f) shows splash screen.

3. Experimental Evaluations

In this section, we evaluate the performance of our proposed approach using the MSCOCO and VizWiz Captions datasets for our experimental assessments.

3.1. Dataset and Performance Metrics

The proposed approach was tested with a dataset consisting of images, each accompanied by five corresponding reference captions. In this study, the approach is trained with two commonly used datasets, namely MSCOCO and VizWiz Captions. The VizWiz Captions dataset (Gurari et al., 2020) consists of 23,431 training, 7,750 validation, and 8,000 test images taken by visually impaired individuals, each paired with five reference captions. Similarly, the MSCOCO Captions dataset (Lin et al., 2014) contains 118,287 training and 5,000 validation images, each annotated with a minimum of five reference captions. Experimental evaluations are conducted using the performance metrics BLEU-n ($n = 1, 2, 3, 4$) (Papineni et al., 2002), ROUGE-L (Ganesan, 2018), SPICE (Anderson et al., 2016), METEOR (Banerjee & Lavie, 2005), and CIDEr (Vedantam et al., 2015).

BLEU-n is an evaluation metric often employed to assess the performance of machine translation systems. It measures the similarity between two sentences by assigning a score of 0 or 1, which is determined by the extent of word overlap between the generated translation and the reference sentences. In BLEU-n, the parameter "n" denotes the number of words taken into account when calculating the similarity. Likewise, METEOR was developed for machine translation and utilizes the harmonic mean of precision and recall of unigram matches. SPICE parses each reference sentence and evaluates the objects, attributes, and relationships in the generated captions, assessing semantic performance, rather than directly comparing a generated caption to a set of reference sentences for syntactic compatibility. ROUGE-L is a performance metric for text summarization that measures the longest common subsequence between the generated summaries and their references. CIDEr assesses the consensus between a caption and references by leveraging sentence similarity, capturing both grammatical correctness and salient concepts. CIDEr metric exhibits varying degrees of similarity to human judgment and evaluates prominent attributes and relationships in both linguistic and semantic aspects. Therefore, priority is given to CIDEr results when comparing the outcomes. The ranking of the results is determined by a final SCORE, which is calculated as the average of all performance metrics. The calculation of the final SCORE takes place using the average of BLEU scores and other performance metrics.

3.2. Results and Discussion

The effectiveness of the fusion of high-level visual attributes is measured on the MSCOCO and VizWiz Captions validation sets. In the experiments, we employed semantic segmentation, object detection, instance segmentation, and keypoint detection features for image captioning.

Table 1 presents how various visual attributes, such as Instance Segmentation (IS), Keypoint Detection (KD), Object Detection (OD), and Semantic Segmentation (SS), affect the performance of image captioning on the MSCOCO Captions dataset. Furthermore, when pairing IS with KD or KD with OD, the results tend to improve in terms of performance metrics. This indicates that combining different visual attributes enhance the performance of image captioning. When we use all the visual attributes on the MSCOCO dataset, we see the best performance metric scores in CIDEr, BLEU-4, SPICE, and the overall SCORE. This indicates that using the fusion of high-level visual attributes can be quite effective.

Table 2 Sample images from VizWiz (first two columns) and MSCOCO (last two columns) with ground-truth and generated captions





			
Reference Captions			
A can of crushed tomatoes are on a brown surface, the tomatoes read crushed tomatoes on the brand.	Its is a basil leaves container its contains the net weight too.	A man riding a skateboard down the side of a ramp.	A girl in a bathing suit with a pink umbrella.
A can of crushed tomatoes sitting on a beige colored counter.	A green and white plastic condiment bottle containing Basil leaves.	Tree is a male skateboarder that is riding a ramp.	A woman in a floral swim- suit holds a pink umbrella.
A can of crushed tomatoes in puree from price chopper.	Quality issues are too severe to recognize visual content.	A boy is skateboarding down a ramp and catches some air.	A woman posing for the camera, holding a pink, open umbrella and wearing a bright, floral, ruched bathing suit, by a lifeguard stands with lake, green trees, and a blue sky with a few clouds behind.
A Price Chopper branded can of crushed tomatoes.	A bottle of spices in a plastic container laying on a surface.	A young boy skateboarding down a ramp at a skate park.	A woman with an umbrella near the sea.
Image is a can of crushed tomatoes in view.	Some basil leaves in a con- tainer on a counter.	A person on a skateboard jumping on a ramp.	Woman in swimsuit holding parasol on a sunny day.
Generated caption			
A can of diced tomatoes that is unopened	A bowl of green leaves sitting on a table	A man riding a skateboard up the iside of a ramp	A woman standing by a river with a red umbrella

Figure 2 illustrates that IS, OD, and SS visual attributes correlate with one another, while KD differs due to its focus on human body parts, unlike the scenes and events in the MSCOCO Captions dataset. Table 2 shows generated and reference captions for two sample images from the VizWiz Captions dataset. For the first image, the generated caption identifies the object as "a can of diced tomatoes that is unopened.". While this caption captures the primary object, it differs slightly in the description. The reference captions, in contrast, describe the can as containing "crushed" rather than "diced" tomatoes. The generated caption for the second image describes "a bowl of green leaves sitting on a table." This is aligned with a few of the reference captions which mention "basil leaves in a container." However, there are nuances missed by the generated caption, such as the specific container type and its positioning. Notably, one of the reference captions states "Quality issues are too severe to recognize visual content", hinting at potential challenges in image clarity that could influence caption generation. Overall, while the generated captions provide a general understanding of the depicted scenes, they occasionally lack the specificity and accuracy seen in the reference captions.

Table 2 presents sample images from the MSCOCO Captions dataset alongside both reference and generated captions. For the first image, featuring a skateboarder, the generated caption identifies the main action, which is a man riding a skateboard. However, it inaccurately describes the direction, interpreting the skateboarder as going "up" the ramp. However, reference captions state that the man going "down" the ramp. The second image presents a woman with an umbrella near a body of water. The generated caption perceives her standing by a "river" with a "red" umbrella. However, the reference captions describe the umbrella as "pink" and the body of water as either the sea or a lake. In both cases, the generated captions grasp the primary elements of the images, but there are discrepancies in specific details when compared to the reference captions. This indicates that while the image captioning system recognizes the main elements, it could be improved with more attention to detail.

The interface of the application is illustrated across various screenshots in Figure 3. Figure 3 (a) shows the initial startup screen of the application. Once a user logs in, they get to the home screen, as presented in Figure 3 (b). This home screen is equipped with several functional buttons: one for capturing an image, another to toggle the flashlight, and the last to access the personal gallery of the user. Figure 3 (c) shows the caption automatically generated from a captured image. To change the language of the caption, Figure 3 (d) provides a language selection feature. If a user wants to choose an image from the gallery, they would use the gallery icon, with the resulting interface depicted in Figure 3 (e). Once a language is chosen, the translation of the caption appears as demonstrated in Figure 3 (f), corresponding to the language selection from Figure 3 (d).

4. Conclusion

In this paper, we present the fusion of high-level visual attributes under the encoder-decoder framework for accurate image captioning. This fusion aims to capture a richer representation of the image, ensuring more detailed and accurate caption generation. The proposed approach integrates models including Instance Segmentation, Semantic Segmentation, Object Detection, and Keypoint Detection into the encoder, leveraging their pre-trained weights for enhanced performance. On the decoder side, a residual-connected GRU-based model is utilized to generate the corresponding captions. The evaluation of this approach was conducted on the VizWiz and MSCOCO Captions datasets, and when visual attributes were fused it achieved the highest score. Furthermore, a user-friendly Android application called *NObstacle* was developed, which exhibits significant potential in assisting visually impaired individuals with their daily activities. Future research involves the utilization of transformers to enhance the accuracy of captions in terms of performance metrics.

Acknowledgment

This research was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) British Council (The Newton Katip Celebi Fund Institutional Links, Turkey UK project: 120N995) and by the scientific research projects coordination unit of Izmir Katip Celebi University (project no: 2021-ÖDL-MÜMF-0006, & 2022-TYL-FEBE-0012).

References

- Akosman, Ş. A., Öktem, M., Moral, Ö. T., & Kılıç, V. (2021). Deep Learning-based Semantic Segmentation for Crack Detection on Marbles. 29th Signal Processing and Communications Applications Conference (SIU),
- Amit, Y., Felzenszwalb, P., & Girshick, R. (2020). Object detection. *Computer Vision: A Reference Guide*, 1-9.
- Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. Computer Vision–ECCV: 14th European Conference, Amsterdam, The Netherlands, October 11-14, Proceedings, Part V 14,
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization,
- Baran, M., Moral, Ö. T., & Kılıç, V. (2021). Akıllı telefonlar için birleştirme modeli tabanlı görüntü altyazılama. *Avrupa Bilim ve Teknoloji Dergisi*(26), 191-196.
- Barroso-Laguna, A., Riba, E., Ponsa, D., & Mikolajczyk, K. (2019). Key. net: Keypoint detection by handcrafted and learned cnn filters. Proceedings of the IEEE/CVF international conference on computer vision,
- Betül, U., Çaylı, Ö., Kılıç, V., & Onan, A. (2022). Resnet based deep gated recurrent unit for image captioning on smartphone. *Avrupa Bilim ve Teknoloji Dergisi*(35), 610-615.
- Aydın, S., Çaylı, Ö., Kılıç, V., & Onan, A. (2022). Sequence-to-sequence video captioning with residual connected gated recurrent units. *Avrupa Bilim ve Teknoloji Dergisi*(35), 380-386.
- Çaylı, Ö., Makav, B., Kılıç, V., & Onan, A. (2021). Mobile application based automatic caption generation for visually impaired. Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS Conference, Istanbul, Turkey, July 21-23,
- Chang, S.-F. (1995). Compressed-domain techniques for image/video indexing and manipulation. Proceedings., International Conference on Image Processing,
- Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., & Luo, J. (2018). "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention. Proceedings of the european conference on computer vision (ECCV),
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Çaylı, Ö., Kılıç, V., Onan, A., & Wang, W. (2022). Auxiliary classifier based residual rnn for image captioning. 30th European Signal Processing Conference (EUSIPCO),
- Çaylı, Ö., Liu, X., Kılıç, V., & Wang, W. (2023). Knowledge Distillation for Efficient Audio-Visual Video Captioning. *arXiv preprint arXiv:2306.09947*.
- Deselaers, T., Keysers, D., & Ney, H. (2004). Features for image retrieval: A quantitative comparison. Pattern Recognition: 26th DAGM Symposium, Tübingen, Germany, August 30-September 1. Proceedings 26,
- Doğan, V., Isık, T., Kılıç, V., & Horzum, N. (2022). A field-deployable water quality monitoring with machine learning-based smartphone colorimetry. *Analytical Methods*, 14(35), 3458-3466.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. Computer Vision–ECCV: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, Proceedings, Part IV 11,
- Fetiler, B., Çaylı, Ö., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). Video captioning based on multi-layer gated recurrent unit for smartphones. *Avrupa Bilim ve Teknoloji Dergisi*(32), 221-226.
- Ganesan, K. (2018). Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

- Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7, 87-93.
- Gurari, D., Zhao, Y., Zhang, M., & Bhattacharya, N. (2020). Captioning images taken by people who are blind. Computer Vision–ECCV: 16th European Conference, Glasgow, UK, August 23–28, Proceedings, Part XVII 16,
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Proceedings of the IEEE international conference on computer vision,
- Ibarra, F. F., Kardan, O., Hunter, M. R., Kotabe, H. P., Meyer, F. A., & Berman, M. G. (2017). Image feature types and their predictions of aesthetic preference and naturalness. *Frontiers in Psychology*, 8, 632.
- Jiang, W., Ma, L., Chen, X., Zhang, H., & Liu, W. (2018). Learning to guide decoding for image captioning. Proceedings of the AAAI Conference on Artificial Intelligence,
- Keskin, R., Çaylı, Ö., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). A benchmark for feature-injection architectures in image captioning. *Avrupa Bilim ve Teknoloji Dergisi*(31), 461-468.
- Kılıç, V. (2021). Deep gated recurrent unit for smartphone-based image captioning. *Sakarya University Journal of Computer and Information Sciences*, 4(2), 181-191.
- Kılıç, V., Mercan, Ö. B., Tetik, M., Kap, Ö., & Horzum, N. (2022). Non-enzymatic colorimetric glucose detection based on Au/Ag nanoparticles using smartphone and machine learning. *Analytical Sciences*, 38(2), 347-358.
- Kılıç, V., Zhong, X., Barnard, M., Wang, W., & Kittler, J. (2014). Audio-visual tracking of a variable number of speakers with a random finite set approach. 17th International Conference on Information Fusion (FUSION),
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Computer Vision–ECCV: 13th European Conference, Zurich, Switzerland, September 6-12, Proceedings, Part V 13,
- Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Makav, B., & Kılıç, V. (2019). Smartphone-based image captioning for visually and hearing impaired. 11th international conference on electrical and electronics engineering (ELECO),
- Mercan, Ö. B., Doğan, V., & Kılıç, V. (2020). Time Series Analysis based Machine Learning Classification for Blood Sugar Levels. Medical Technologies Congress (TIPTEKNO),
- Mercan, Ö. B., & Kılıç, V. (2020). Deep learning based colorimetric classification of glucose with au-ag nanoparticles using smartphone. Medical Technologies Congress (TIPTEKNO),
- Moral, Ö. T., Kılıç, V., Onan, A., & Wang, W. (2022). Automated Image Captioning with Multi-layer Gated Recurrent Unit. 30th European Signal Processing Conference (EUSIPCO),
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics,
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Pu, B., Liu, Y., Zhu, N., Li, K., & Li, K. (2020). ED-ACNN: Novel attention convolutional neural network based on encoder–decoder framework for human traffic prediction. *Applied Soft Computing*, 97, 106688.
- Sayraci, B., Ağralı, M., & Kılıç, V. (2023). Artificial Intelligence Based Instance-Aware Semantic Lobe Segmentation on Chest Computed Tomography Images. *Avrupa Bilim ve Teknoloji Dergisi*(46), 109-115.
- Sun, X., Xiao, B., Wei, F., Liang, S., & Wei, Y. (2018). Integral human pose regression. Proceedings of the European conference on computer vision (ECCV),
- Tahir, H., Iftikhar, A., & Mumraiz, M. (2021). Forecasting COVID-19 via registration slips of patients using resnet-101 and performance analysis and comparison of prediction for COVID-19 using faster r-cnn, mask r-cnn, and resnet-50. International conference on advances in electrical, computing, communication and sustainable technologies (ICAECT),
- Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., & Cottrell, G. (2018). Understanding convolution for semantic segmentation. IEEE winter conference on applications of computer vision (WACV),
- Xi, D., Qin, Y., Luo, J., Pu, H., & Wang, Z. (2021). Multipath fusion mask R-CNN with double attention and its application into gear pitting detection. *IEEE Transactions on Instrumentation and Measurement*, 70, 1-11.
- Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., & Li, C. (2020). An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Transactions on Image Processing*, 29, 9627-9640.
- You, Q., Jin, H., & Luo, J. (2018). Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*.
- Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12), 4467-4480.