Avrupa Bilim ve Teknoloji Dergisi
Sayı 53, S. 58-63, Ocak 2024

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 53, pp. 58-63, January 2024

Research Article

Leveraging Pre-trained 3D-CNNs for Video Captioning

Bengü Fetiler[1*], Özkan Çaylı[1], Volkan Kılıç[1]

[1] İzmir Katip Çelebi University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, İzmir, Turkey, (ORCID: 0000-0002-2761-7751, 0000-0002-3389- 3867, 0000-0002-3164-1981), y220207008@ogr.ikc.edu.tr, ozkan.cayli@ikcu.edu.tr, volkan.kilic@ikcu.edu.tr

(İlk Geliş Tarihi 6 Ekim 2023 ve Kabul Tarihi 19 Kasım 2023)

(**DOI:** 10.5281/zenodo.10623749)

ATIF/REFERENCE: Fetiler, B., Çaylı, Ö., & Kılıç, V. (2024). Leveraging Pre-trained 3D-CNNs for Video Captioning. *European Journal of Science and Technology*, (53), 58-63.

Abstract

Video captioning is a visual understanding task that aims to generate grammatically and semantically accurate descriptions. One of the main challenges in video captioning is capturing the complex dynamics present in videos. This study addresses this challenge by leveraging pre-trained 3D Convolutional Neural Networks (3D-CNNs). These networks are particularly effective at modeling such dynamics, enhancing video contextual understanding. We evaluated the approach on the Microsoft Research Video Description (MSVD) dataset, with commonly utilized performance metrics in video captioning including CIDEr, BLEU-1 through BLEU-4, ROUGE-L, METEOR, and SPICE. The results show significant improvements across all these metrics, proving the advantage of pre-trained 3D-CNNs in enhancing video captioning accuracy.

Keywords: Video Captioning, Video-Language Multimodal Learning, Motion Features.

Video Altyazılama için Önceden Eğitilmiş 3B-CNN'lerden Yararlanma

Öz

Video altyazılama, hem dilbilgisel hem de anlamsal olarak doğru açıklamalar oluşturmayı amaçlayan bir görsel anlama görevidir. Video altyazılamadaki ana zorluklardan biri, videolardaki karmaşık dinamikleri yakalamaktır. Bu çalışma bu zorluğu aşmak için önceden eğitilmiş 3B Evrişimli Sinir Ağlarını (3D-CNNs) kullanmaktadır. Bu ağlar bu tür dinamikleri modellemede özellikle etkilidir, böylece videoların bağlamsal anlayışını artırır. Önerilen yaklaşım, video altyazılama için yaygın olarak tanınan bir ölçüt olan Microsoft Araştırma Video Açıklama (MSVD) veri seti üzerinde değerlendirildi. Performansı değerlendirmek için BLEU-1'den BLEU-4'e, CIDEr, ROUGE-L, METEOR ve SPICE de dahil olmak üzere standart metrikler kullandık. Sonuçlar, tüm bu metriklerde önemli iyileşmeler göstererek, önceden eğitilmiş 3D-CNN'lerin video altyazılama doğruluğunu artırdığını vurgulamaktadır.

Anahtar Kelimeler: Video Altyazılama, Video-Dil Multimodal Öğrenme, Hareket Nitelikleri.

* Corresponding Author: y220207008@ogr.ikc.edu.tr

1. Introduction

Video captioning is a task that involves generating descriptions for video frames by leveraging techniques from natural language processing and computer vision fields. These descriptions are expected to be grammatically correct and semantically accurate. Recently, there has been increased attention on video captioning studies due to their potential applications in video understanding, video retrieval, and video caption generation (Çaylı et al., 2023; Gan et al., 2016; Guo et al., 2016; Shen et al., 2013).

Earlier studies in captioning have explored various approaches, including template-based, retrieval-based, and deep learning- based. One template-based approach uses a predefined template to translate semantic representation into a caption (Venugopalan et al., 2014). The retrieval-based approach employs a compositional semantics language model that breaks down video descriptions into subjects, verbs, and objects. These elements are then transformed into word vectors, effectively capturing the meaning of the content (Guadarrama et al., 2013).

Recently, deep learning-based approaches have emerged as valuable tools for generating more accurate captions (Aydın et al., 2022; Baran et al., 2021; Çaylı et al., 2022; Çaylı et al., 2021; Fetiler et al., 2021; Keskin, Çaylı, et al., 2021; Keskin, Moral, et al., 2021; Kılcı et al., 2023; Kılıç, 2021; Makav & Kılıç, 2019; Uslu et al., 2022). These approaches leverage deep learning to manage the complexity of videos, including diverse objects, scenes, and actions. Various deep learning-based encoder-decoder architectures have been proposed. These architectures typically combine convolutional neural networks (CNNs) to extract features and recurrent neural networks (RNNs) for caption generation (Akosman et al., 2021; Amaresh & Chitrakala, 2019; Doğan et al., 2022; Kılıç et al., 2022; Kılıç et al., 2014; Koca & Kılıç, 2023; Koca et al., 2023; Mercan et al., 2020; Mercan & Kılıç, 2020; Moral et al., 2022; Palaz et al., 2021; Sayracı et al., 2023). There are various CNN architectures commonly employed in the encoder for feature extraction from video frames to feed RNN-based decoders (Chollet, 2017; Doğan et al., 2024; Szegedy et al., 2016; Targ et al., 2016). However, conventional RNNs encounter challenges such as vanishing and exploding gradient issues, limiting their ability to process long input sequences due to short-term memory. Two types of RNNs have been proposed to address these challenges: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). LSTM networks introduce three gates: the input gate, the forget gate, and the output gate. These gates, along with two states known as the hidden state and memory cells, enable LSTMs to capture long-term dependencies in sequences effectively. On the other hand, GRU networks consist of a hidden state and two gates: the update and the reset gate. GRUs can dynamically determine, by utilizing these gates, the amount of information to retain from previous time steps and update their hidden state accordingly. This enables GRUs to model dependencies in sequences with varying lengths.

A video captioning approach that utilizes the encoder-decoder architecture incorporates a hierarchical recurrent neural encoder (HRNE) with a two-layer LSTM (P. Pan et al., 2016a). The HRNE extracts temporal features from video frames, which serve as input for the LSTM-based decoder that generates captions. The LSTM hidden state and memory cell are carried forward to the next step, except when a new video time boundary is detected.

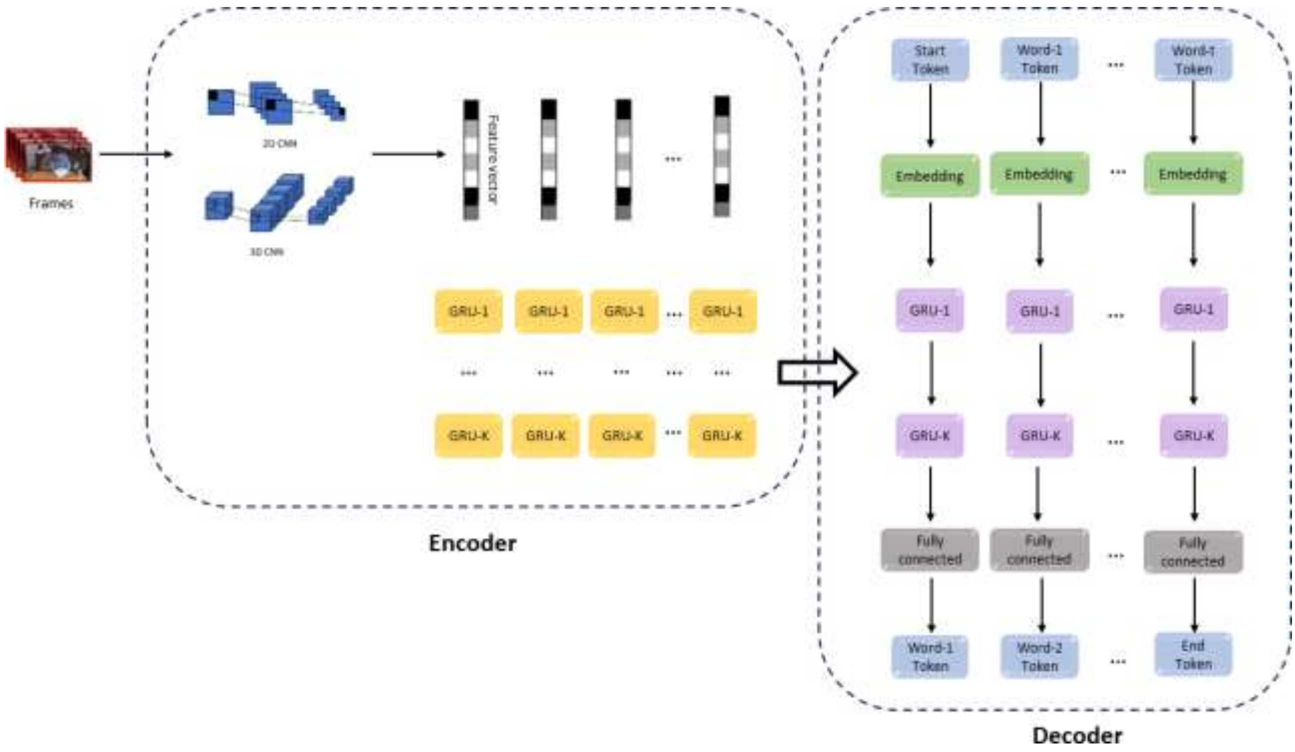

Figure 1- Proposed Approach

Table 1. Comparison of Different 3D-CNN Architectures with Inception-v3

Multi-layer GRUs	# of Layers	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	SPICE
S3D+Inception-v3	1	0.860	0.494	0.588	0.678	0.802	**0.712**	**0.350**	0.058
	2	**0.863**	0.493	0.588	0.685	0.807	**0.712**	**0.350**	0.060
	4	0.789	0.492	**0.593**	**0.693**	**0.814**	0.702	0.335	**0.064**
R3D+Inception-v3	1	0.770	0.442	0.547	0.618	0.780	0.692	0.330	0.054
	2	0.809	0.453	0.550	0.654	0.784	0.700	0.330	0.055
	4	0.850	**0.502**	0.585	0.684	0.808	0.711	0.339	0.061
P3D+Inception-v3	1	0.785	0.462	0.561	0.651	0.774	0.700	0.329	0.054
	2	0.822	0.478	0.576	0.672	0.793	0.708	0.337	0.058
	4	0.808	0.477	0.584	0.684	0.805	0.704	0.330	0.063
MVIT+Inception-v3	1	0.803	0.458	0.561	0.661	0.786	0.700	0.330	0.055
	2	0.716	0.465	0.562	0.653	0.782	0.699	0.333	0.056
	4	0.820	0.482	0.582	0.680	0.801	0.708	0.333	0.058
Inception-v3	4	0.715	0.491	0.591	0.692	0.813	0.701	0.334	0.063
S3D	4	0.788	0.491	0.592	0.691	0.813	0.701	0.335	0.063
R3D	4	0.513	0.370	0.471	0.574	0.720	0.651	0.288	0.043
P3D	4	0.230	0.270	0.373	0.488	0.670	0.621	0.240	0.038
MVIT	4	0.181	0.330	0.490	0.601	0.742	0.611	0.240	0.040

The Sequence-to-Sequence Video-to-Text (S2VT) approach was proposed for video captioning to capture the temporal structure of videos and represent them as fixed-length vectors. This S2VT approach employs LSTMs in both its encoder and decoder, facilitating the encoding of the temporal structure of video and the generation of captions (Venugopalan et al., 2015).

In this paper, we propose a video captioning approach with a combination of two-dimensional (2D) and 3D-CNN architectures and multi-layer GRU to extract features of the videos on the encoder side. Inception-v3 as 2D-CNN is employed to extract appearance features from video frames, whereas S3D, R3D, P3D, and MVIT as 3D-CNNs are utilized for the motion features. Then, a multi-layer GRU is employed to preserve the semantic information of the video and leverage contextual information more effectively. On the decoder side, a multi-layer GRU is utilized to generate more accurate captions by leveraging its ability to compute complex representations. Experimental results are obtained on the MSVD dataset using various evaluation metrics, including BLEU-n (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and SPICE (Anderson et al., 2016). These metrics are used to measure the accuracy of the proposed approach on captioning performance and to compare with state-of-the-art approaches.

The rest of this paper is structured as follows: Section 2 introduces the 2D and 3D-based sequence-to-sequence approach for video captioning. In Section 3, we describe the dataset, performance metrics, and results achieved by our proposed approach. Section 4 provides the conclusion and outlines future research directions.

2. Proposed Approach

In this section, we introduce our proposed approach as shown in Figure 1 for video captioning based on sequence-to-sequence learning which utilizes pre-trained 3D-CNNs.

The proposed video captioning approach is employed under the encoder-decoder framework. In this framework, the encoder extracts visual attributes from videos. These extracted attributes are then fed into the decoder, which generates descriptive captions detailing events and scenes corresponding to relevant parts of the video.

For each iteration, the multi-layer GRU of the encoder receives the updated hidden state from the previous iteration until it reaches the last feature vector. The final hidden state of the multi-layer GRU in the encoder is then passed to the decoder for caption generation. The video decoder consists of an embedding layer, a multi-layer GRU, and a fully connected layer. Caption generation begins with a predefined start token at $t = 0$ and continues for a variable length T. The embedding layer transforms each token into a meaningful latent vector containing linguistic features. The latent vector is then provided as input to the first GRU layer. The output from this layer is then transferred to the following layer. This procedure is carried out K times, with K denoting the total count of GRU layers. The output of the multi-layer GRU is then directed into a fully connected layer, which calculates the prediction probabilities and generates the subsequent word in the caption. The fully connected layer generates the token for the first word (word-1), which will be used in the following step. This word generation procedure continues for T iterations until the end token is reached.

All generated tokens are converted into their corresponding words to form a caption. To evaluate the impact on captioning performance, we varied the number of GRU layers, testing configurations with 1, 2, and 4 layers on both the encoder and the decoder sides.

Table 2. Performance Comparison of the Proposed Approach and State-of-the-Art Architectures on the MSVD Dataset

	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	METEOR
(P. Pan et al., 2016b)	-	0.438	0.551	0.663	0.792	0.331
(Baraldi et al., 2017)	0.635	0.425	-	-	-	0.324
(Yao et al., 2015)	0.517	0.419	0.526	0.647	0.800	0.296
(Yu et al., 2016)	0.658	0.499	**0.604**	**0.704**	**0.815**	0.326
(Y. Pan et al., 2016)	-	0.453	0.554	0.660	0.788	0.310
Proposed S3D with 2-layer GRU	**0.863**	**0.493**	0.588	0.685	0.807	**0.350**

3. Experimental Evaluations

3.1. Dataset

Various datasets, M-VAD (Torabi et al., 2015), MPII-MD (Rohrbach et al., 2015), MSR-VTT (Xu et al., 2016), and MSVD (Chen & Dolan, 2011), have been employed to evaluate the performance of various video captioning approaches. M-VAD comprises 48,986 videos extracted from 92 movies, 38,949 training, 4,888 validation, and 5,149 test videos. Each video in M-VAD is annotated with a single caption. The MPII-MD dataset is a large-scale collection of approximately 68,337 videos, with each video possessing one reference caption. MSR-VTT contains 10,000 videos with diverse content, such as news and sports. Each video in this dataset is annotated with 20 reference captions. The MSVD dataset consists of 1,200 training videos, 100 validation videos, and 670 test videos sourced from YouTube. Each video in MSVD is associated with 40 captions. We chose the MSVD dataset for the evaluation of our proposed video captioning approach due to its extensive reference captions.

3.2. Evaluation Metrics

The performance of the video captioning approaches is evaluated using several metrics, including BLEU-n (n = 1, 2, 3, 4), METEOR, ROUGE-L, SPICE, and CIDEr. BLEU-n measures the similarity between a machine-generated caption and reference captions. BLEU-n considers n-grams (contiguous sequences of n words) to evaluate the quality of the generated caption. METEOR evaluates the overall quality of the generated caption by considering various aspects such as precision, recall, and alignment with the reference captions. ROUGE-L measures the similarity between the generated caption and reference captions based on the longest common subsequence of words. SPICE is designed explicitly for captioning tasks and evaluates the semantic suggestive content of the generated and reference captions. CIDEr, also designed for captioning, calculates the average cosine similarity between the generated and reference captions. CIDEr is often used to sort the results in image and video captioning tasks due to its better correlation with human judgment than BLEU-n, METEOR, SPICE, and ROUGE-L. For our evaluation, we prioritize the CIDEr metric to sort results, as it aligns more closely with human judgment compared to the other metrics.

3.3. Results and Discussion

Table 1 comprehensively evaluates various 3D-CNN architectures paired with Inception-v3, using CIDEr, BLEU (1-4), ROUGE-L, METEOR, and SPICE metrics. The S3D+Inception-v3 Multi-layer GRU with 2 layers demonstrated superior performance, yielding a CIDEr score of 0.863, which indicates its enhanced ability to generate accurate descriptions of videos aligned with human annotations. Furthermore, it showed consistent performance across the BLEU-3, BLEU-2, and BLEU-1 metrics. The four-layer S3D+Inception-v3 Multi-layer GRU outperformed in terms of the SPICE metric, highlighting its proficiency in evaluating semantic content. Moreover, the R3D+Inception-v3 Multi-layer GRU with 4 layers achieved a remarkable BLEU-4 score of 0.502.

Table 2 benchmarks the proposed S3D with a 2-layer GRU against state-of-the-art approaches on the MSVD dataset. Remarkably, the proposed approach achieves the highest CIDEr (0.863) and METEOR (0.350) scores, indicating enhanced video description quality.

Although our approach excels in BLEU-4 (0.493), indicating relevant and coherent long caption generation, it is outperformed by (Yu et al., 2016) in the metrics BLEU-1 to BLEU-3. This demonstrates that their method generates short captions more accurately. The results emphasize the advanced semantic caption generation of the proposed approach while comparing the competitive domain of video captioning architectures.

4. Conclusion

In this study, a video captioning approach has been developed under the encoder-decoder-based sequence-to-sequence approach. Different 2D and 3D-CNN architectures were used to extract the features of the video frames, and a multi-layer GRU was used to process the features and generate the video caption. The evaluations in the MSVD dataset show that the proposed approach improves the accuracy of 3D-CNN architectures in generating meaningful captions. We plan to explore ensembles of 3D-CNN architectures in our future study. Additionally, an evaluation of the feature extraction and representation capabilities of these architectures will be conducted to provide insights into their strengths and weaknesses.

5. Acknowledge

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. 2021 29th Signal Processing and Communications Applications Conference (SIU),

Amaresh, M., & Chitrakala, S. (2019). Video captioning using deep learning: an overview of methods, datasets and metrics. 2019 International Conference on Communication and Signal Processing (ICCSP),

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. European Conference on Computer Vision (ECCV),

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.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization,

Baraldi, L., Grana, C., & Cucchiara, R. (2017). Hierarchical boundary-aware neural encoder for video captioning. Conference on Computer Vision and Pattern Recognition (CVPR),

Baran, M., Moral, Ö. T., & Kılıç, V. (2021). Akıllı telefonlar için birleştirme modeli tabanlı görüntü altyazılama. *Avrupa Bilim ve Teknoloji Dergisi*(26), 191-196.

Chen, D., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies,

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition,

Çaylı, Ö., Kılıç, V., Onan, A., & Wang, W. (2022). Auxiliary classifier based residual rnn for image captioning. 2022 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*.

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 2020 Conference, Istanbul, Turkey, July 21-23, 2020,

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.

Doğan, V., Evliya, M., Kahyaoglu, L. N., & Kılıç, V. (2024). On-site colorimetric food spoilage monitoring with smartphone embedded machine learning. *Talanta*, *266*, 125021.

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.

Gan, C., Yao, T., Yang, K., Yang, Y., & Mei, T. (2016). You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., & Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. Proceedings of the IEEE international conference on computer vision,

Guo, Z., Gao, L., Song, J., Xu, X., Shao, J., & Shen, H. T. (2016). Attention-based LSTM with semantic consistency for videos captioning. Proceedings of the 24th ACM international conference on Multimedia,

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.

Keskin, R., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). Multi-gru based automated image captioning for smartphones. 2021 29th Signal Processing and Communications Applications Conference (SIU),

Kılcı, M., Çaylı, Ö., & Kılıç, V. (2023). Fusion of High-Level Visual Attributes for Image Captioning. *Avrupa Bilim ve Teknoloji Dergisi*(52), 161-168.

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),

Koca, Ö. A., & Kılıç, V. (2023). Multi-Parametric Glucose Prediction Using Multi-Layer LSTM. *Avrupa Bilim ve Teknoloji Dergisi*(52), 169-175.

Koca, Ö. A., Türköz, A., & Kılıç, V. (2023). Tip 1 Diyabette Çok Katmanlı GRU Tabanlı Glikoz Tahmini. *Avrupa Bilim ve Teknoloji Dergisi*(52), 80-86.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out, Makav, B., & Kılıç, V. (2019). Smartphone-based image captioning for visually and hearing impaired. 2019 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. 2020 Medical Technologies Congress (TIPTEKNO),

Mercan, Ö. B., & Kılıç, V. (2020). Deep learning based colorimetric classification of glucose with au-ag nanoparticles using smartphone. 2020 Medical Technologies Congress (TIPTEKNO),

Moral, Ö. T., Kiliç, V., Onan, A., & Wang, W. (2022). Automated Image Captioning with Multi-layer Gated Recurrent Unit. 2022 30th European Signal Processing Conference (EUSIPCO),

Palaz, Z., Doğan, V., & Kılıç, V. (2021). Smartphone-based Multi-parametric Glucose Prediction using Recurrent Neural Networks. *Avrupa Bilim ve Teknoloji Dergisi*(32), 1168-1174.

Pan, P., Xu, Z., Yang, Y., Wu, F., & Zhuang, Y. (2016a). Hierarchical recurrent neural encoder for video representation with application to captioning. Proceedings of the IEEE conference on computer vision and pattern recognition,

Pan, P., Xu, Z., Yang, Y., Wu, F., & Zhuang, Y. (2016b). Hierarchical recurrent neural encoder for video representation with application to captioning. Conference on computer vision and pattern recognition,

Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. Conference on Computer Vision and Pattern Recognition (CVPR),

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. Annual Meeting of the Association for Computational Linguistics (ACL),

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. Proceedings of the IEEE conference on computer vision and pattern recognition,

Sayracı, 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.

Shen, F., Shen, C., Shi, Q., Van Den Hengel, A., & Tang, Z. (2013). Inductive hashing on manifolds. Proceedings of the IEEE conference on computer vision and pattern recognition,

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition,

Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Torabi, A., Pal, C., Larochelle, H., & Courville, A. (2015). Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*.

Uslu, B., 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.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. Conference on Computer Vision and Pattern Recognition (CVPR),

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence-video to text. Proceedings of the IEEE international conference on computer vision,

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. Proceedings of the IEEE conference on computer vision and pattern recognition,

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. International Conference on Computer Vision (ICCV),

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. Conference on Computer Vision and Pattern Recognition (CVPR),