Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 32, S. 221-226, Aralık 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 32, pp. 221-226, December 2021
Copyright © 2021 EJOSAT
Research Article

Video Captioning Based on Multi-layer Gated Recurrent Unit for Smartphones

Bengü Fetiler[1], Özkan Çaylı[1], Özge Taylan Moral[1*], Volkan Kılıç[1], Aytuğ Onan[2]

[1] İzmir Katip Celebi University, Faculty of Engineering and Architecture, Department of Electrical and Electronics, İzmir, Turkey, (ORCID: 0000-0002-2761-7751, 0000-0002-3389-3867, 0000-0003-0482-267X, 0000-0002-3164-1981), 160403027@ogr.ikcu.edu.tr, y200207004@ogr.ikcu.edu.tr, ozgetaylan.moral@ikcu.edu.tr, volkan.kilic@ikcu.edu.tr

[2] İzmir Katip Celebi University, Faculty of Engineering and Architecture, Department of Computer Engineering, İzmir, Turkey, (ORCID: 0000-0002-9434-5880), aytug.onan@ikcu.edu.tr

(International Conference on Design, Research and Development (RDCONF) 2021 - 15 – 18 December 2021)

(**DOI:** 10.31590/ejosat.1039242)

ATIF/REFERENCE: Fetiler, B., Çaylı, Ö., Moral, Ö.T., Kılıç, V., & Onan, A. (2021). Video Captioning Based on Multi-layer Gated Recurrent Units for Smartphones. *European Journal of Science and Technology*, (32), 221-226.

Abstract

Video captioning is the visual understanding process to generate grammatically and semantically meaningful descriptions that are of interest in the fields of computer vision (CV) and natural language processing (NLP). Recent advances in the computing power of the mobile platform have led to many video captioning applications that use CV and NLP techniques. These video captioning applications mainly depend on the encoder-decoder approach running with the internet connection, which employs convolutional neural networks (CNNs) on the encoder and recurrent neural networks (RNNs) on the decoder. However, this approach is not powerful enough to get accurate captioning results, and fast response due to online data transfer. In this paper, therefore, the encoder-decoder approach has been extended with a sequence-to-sequence model under a multi-layer gated recurrent unit (GRU) to generate a semantically more coherent caption. Visual information from image features of each video frame is extracted with ResNet-101 CNN in the encoder to feed the multi-layer GRU based decoder for caption generation. The proposed approach has been compared with the state-of-the-art approaches using experiments on the MSVD dataset under eight performance metrics. In addition, the proposed approach is embedded into our custom-designed Android application, called *WeCap*, capable of faster caption generation without an internet connection.

Keywords: Convolutional Neural Network, Gated Recurrent Units, Natural Language Processing, Video Captioning, Android Application.

Akıllı Telefonlar için Çok Katmanlı Kapılı Tekrarlayan Birim Tabanlı Video Altyazılama

Öz

Video altyazılama, bilgisayarlı görü (CV) ve doğal dil işleme (NLP) alanlarında ilgi çeken dilbilgisel ve anlamsal olarak anlamlı tanımlar oluşturan bir görsel anlama işlemidir. Mobil platformun hesaplama gücündeki son gelişmeler, CV ve NLP tekniklerini kullanan birçok video altyazılama uygulamasının önünü açmıştır. Bu video altyazılama uygulamaları, çoğunlukla, kodlayıcı üzerinde evrişimli sinir ağları (CNN'ler) ve kod çözücü üzerinde tekrarlayan sinir ağları (RNN'ler) kullanan internet bağlantısıyla çalışan kodlayıcı-kod çözücü yaklaşımına bağlıdır. Ancak, bu yaklaşım çevrimiçi veri aktarımından dolayı doğru altyazı sonuçları ve hızlı yanıt alma açısından yeterince güçlü değildir. Bu nedenle, bu bildiride, kodlayıcı-kod çözücü yaklaşımı anlamsal olarak daha uyumlu altyazı oluşturmak için çok katmanlı kapılı tekrarlayan birim (GRU) altında diziden dizeye yaklaşımı ile genişletilmiştir. Her video karesinin görüntü özelliklerinden görsel bilgiler, altyazı oluşturma amacıyla çok katmanlı GRU tabanlı kod çözücüyü beslemek için kodlayıcıdaki ResNet-101 CNN ile çıkarılır. Önerilen yaklaşım, sekiz performans metriği altında MSVD veri kümesi üzerinde deneyler kullanılarak gelişmiş yaklaşımlarla karşılaştırılmıştır. Ayrıca, önerilen yaklaşım internet bağlantısı olmadan daha hızlı altyazı üretme yeteneğine sahip, *WeCap* adlı, özel tasarlanmış Android uygulamamıza gömülmüştür.

Anahtar Kelimeler: Evrişimsel sinir ağı, Kapılı Tekrarlayan Birim, Doğal Dil İşleme, Video Altyazılama, Android Uygulama.

[1] Corresponding Author: ozgetaylan.moral@ikcu.edu.tr

1. Introduction

Video captioning aims to generate grammatically correct and human-readable descriptions of video frames using computer vision and natural language processing techniques. Video captioning has recently received increased interest due to its potential applications, such as video understanding (Gan, Yao, Yang, Yang, & Mei, 2016), video retrieval (Shen, Shen, Shi, Van Den Hengel, & Tang, 2013), and automatic video caption generation (Guo et al., 2016).

Earlier studies on captioning include the template-based approach, which uses a template to translate the semantic representation to a caption (Venugopalan et al., 2014). A compositional semantics language model splits a video description into subjects, verbs and objects, and then converts them into word vectors to capture the meaning of the content (R. Xu, Xiong, Chen, & Corso, 2015). The relationships between subjects, verbs, and objects to capture semantic hierarchies of the video content were studied in (Guadarrama et al., 2013). Deep learning based alternative approaches have recently emerged as a useful tool for generating more accurate captions due to their abilities in dealing with the complexity of videos like diverse objects, scenes, or actions (Amaresh & Chitrakala, 2019; Çaylı, Makav, Kılıç, & Onan, 2020; Keskin, Moral, Kılıç, & Onan, 2021).

Deep learning based encoder-decoder architectures combine convolutional neural networks (CNNs) and recurrent neural networks to design encoder and decoder, respectively, for feature extraction and caption generation. There are several CNN architectures such as Inception-v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), Xception (Chollet, 2017), and ResNet (Targ, Almeida, & Lyman, 2016) that are commonly used in encoder design to extract features of video frames to feed the RNN-based decoders. However, conventional RNNs suffer from vanishing and exploding gradient problems, which prevent handling long input sequences due to short-term memory. Long short-term memory (LSTM) and gated recurrent unit (GRU) networks are proposed to solve these problems with the addition of cells and gates to RNN. LSTM has three gates as input, forget and output while two vector states as hidden and memory cells. GRU consists of a hidden state and two gates: update gate and reset gate. It computes the hidden state and output vector using the input vector with the variable length.

A video captioning model designed using the encoder-decoder architecture utilizes a hierarchical recurrent neural encoder (HRNE) using a two-layer LSTM (P. Pan, Xu, Yang, Wu, & Zhuang, 2016). Temporal features of video frames are extracted with two-layer LSTM as the input of the LSTM based decoder to generate a caption. A hierarchical encoder using a time boundary-aware LSTM cell is proposed to predict the change of actions and events in the video, namely the video time boundary (Baraldi, Grana, & Cucchiara, 2017). The hidden state and memory cell of the LSTM are transferred to the next step unless a new video time boundary is detected and rebooted when the change is detected. The single-layer GRU generates video captions using the outputs of the hierarchical encoder. Contrary to (Baraldi et al., 2017), LSTM with visual-semantic embedding (LSTM-E) model uses CNNs on the encoder side (Y. Pan, Mei, Yao, Li, & Rui, 2016). Accordingly, two-dimensional (2D) CNN is employed to extract spatial features from video frames, whereas three-dimensional (3D) CNN is utilized for temporal features. LSTM-E generates video captions considering the semantic relationship between words. An end-to-end model extracts features of video frames using the average pooling of the 2D-CNN architectures and decodes video features as a sequence of words by LSTM to generate video captions (Venugopalan et al., 2014). Sequence-to-sequence video to text (S2VT) model is proposed to utilize LSTM for encoding the temporal structure of videos to fixed-length vectors both in the encoder and the decoder (Venugopalan et al., 2015).

In this paper, we propose a video captioning system with a combination of ResNet-101 CNN architecture (He, Zhang, Ren, & Sun, 2016) and multi-layer GRU to extract features of the videos on the encoder side. In order to define the visual information, the video is split into frames, and features are extracted using ResNet-101 CNN architecture. Then a multi-layer GRU is utilized to preserve the semantic information of the video and benefit more contextual information. On the decoder side, multi-layer GRU is employed to generate more accurate captions using the ability of computing more complex representations in learning sequential data. Experimental results are obtained on MSVD dataset with BLEU-n (Papineni, Roukos, Ward, & Zhu, 2002), CIDEr (Vedantam, Lawrence Zitnick, & Parikh, 2015), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and SPICE (Anderson, Fernando, Johnson, & Gould, 2016). The effect of multi-layer on captioning performance is measured with these metrics, and the proposed approach is also compared with the state-of-art approaches under the same metrics.

The rest of this paper is organized as follows: Section 2 introduces an encoder-decoder based sequence-to-sequence approach for video captioning and our custom-designed Android application. Section 3 presents the dataset, performance metrics and the results of the proposed approach. Concluding remarks with future research directions are given in Section 4.

2. Proposed Video Captioning Approach

This section presents a sequence-to-sequence approach based on multi-layer GRU for video captioning and our custom-designed Android application, *WeCap*, which runs the proposed system in offline mode.

2.1. Encoder-decoder based Sequence-to-Sequence Approach

In the sequence-to-sequence approach, the video encoder utilizes architectures based on CNN and RNN to extract features and process these features sequentially, while the video decoder employs RNN to generate meaningful captions word by word. GRU is one of the RNN architectures employed to solve the vanishing gradient problem in our approach. Figure 1 depicts our proposed video captioning approach, consisting of CNN and GRU in the encoder, while embedding, GRUs and fully connected layer are on the decoder side. As a CNN architecture, ResNet-101, composed of 101 layers, is preferred in this encoder design due to its capability to solve the degradation problem using a deep residual learning framework. In our approach, the input video is split into frames and is fed to the video encoder. The ResNet-101 extracts all features of frames as a 2048-element vector. Then, the feature vector of each frame is fed into the input vector of the multi-layer GRU to represent features as a single vector at a time for the video decoder. Each initial hidden state is initialized at the beginning. For the successive iterations, the multi-layer GRU of the encoder is fed by the updated hidden state from the previous iteration until reaching to last feature vectors. The last hidden

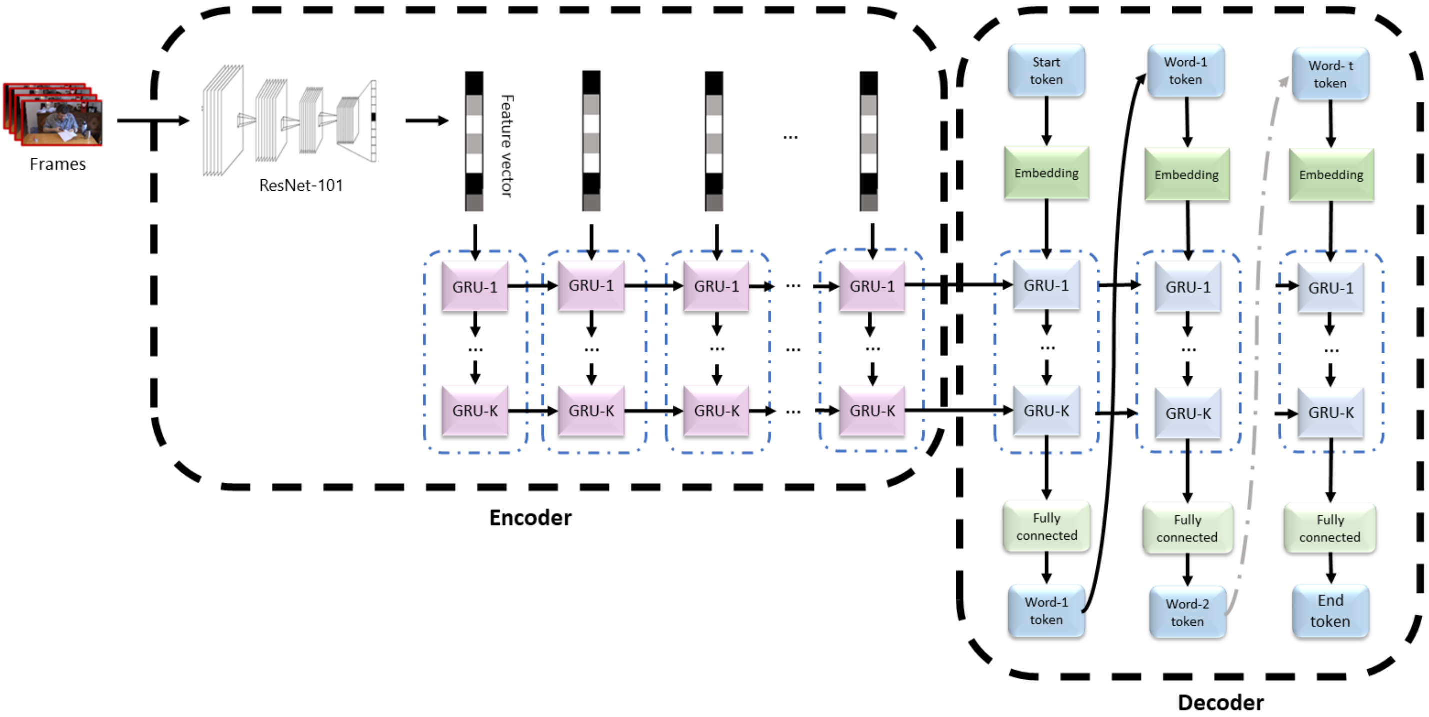


Figure 1. Proposed video captioning approach

state of the multi-layer GRU of the encoder feeds to the decoder for caption generation. The video decoder consists of an embedding, multi-layer GRU, and a fully connected layer. Caption generation starts with the predefined start-token at $t = 1$ for the variable length t. The embedding layer converts each token to the meaningful embedding vector, which includes the linguistic features. The embedding vector is fed to the input vector of the first GRU layer, and the output vector of the first GRU layer is fed to the next layer. This procedure is repeated K-times where K denotes the number of the layer in the GRU. The output of the multi-layer GRU feeds to the fully connected layer to compute the prediction probabilities and generate the following word in the caption. The fully connected layer generates the word-*1* token to utilize in the next step, and the generation continues to t-times to reach the end token. All generated tokens are converted into corresponding words to lead a caption. The layer size of the GRU is increased from a single layer to six layers on both the encoder and the decoder side to see the effects on captioning performance.

2.2. Android Application: *WeCap*

The proposed approach is integrated with our custom-designed Android application, *WeCap*, to offer the benefit of video captioning to non-expert users. In *WeCap*, the encoder-decoder approach is embedded in the smartphone application to reduce the caption generation time, unlike similar image captioning applications (Çaylı et al., 2020; Kılıç, 2021; Makav & Kılıç, 2019), that use cloud connection. A video can be chosen from the gallery by tapping the "gallery" or captured by tapping the "camera" button on the home screen. The user can access these options by scrolling left from the page and with a voice command using the "microphone" button. The embedded encoder-decoder generates captions automatically when the video is chosen. The quantization method has been applied to the embedded encoder-decoder to remove the cloud connection and reduce the size for video captioning. The quantized model is converted to TorchScript and integrated into the application. Caption generation time is significantly reduced in offline mode that extracts features and generates captions without a cloud connection. The generated caption is in English by default and displayed on the screen along with the video. The user can also listen to the generated caption by tapping the "speaker" button. In addition, the user can change the language of the generated caption from the settings, and the Google Cloud Translation API translates the caption into other languages. The interface of the application is illustrated in Figure 2.

3. Experimental Results

In this section, the dataset and performance metrics that are utilized to evaluate the proposed captioning system and the experimental results are presented.

3.1. Dataset and Performance Metrics

To evaluate the performance of the video captioning approaches, several datasets such as M-VAD (Torabi, Pal, Larochelle, & Courville, 2015), MPII-MD (Rohrbach, Rohrbach, Tandon, & Schiele, 2015), MSR-VTT (J. Xu, Mei, Yao, & Rui, 2016), and MSVD (Chen & Dolan, 2011) have been introduced. M-VAD contains 48986 videos from 92 movies consisting of 38949 training, 4888 validation, and 5149 test videos, and also each video is described in a single sentence. The MPII-VD is a large-scale dataset of approximately 68337 videos with one reference caption per video containing audio and movie descriptions. MSR-VTT consists of 10000 videos with different content such as news and sports. Each video is described with 20 reference captions. The MSVD contains 1200 training, 100 validation, and 670 test videos collected from YouTube with 40 captions for each video. In this paper, MSVD is chosen for the evaluation of our proposed video captioning system due to its large reference caption set. In M-VAD and MPII-VD, it is hard to describe a new video as each video is described with a single sentence.

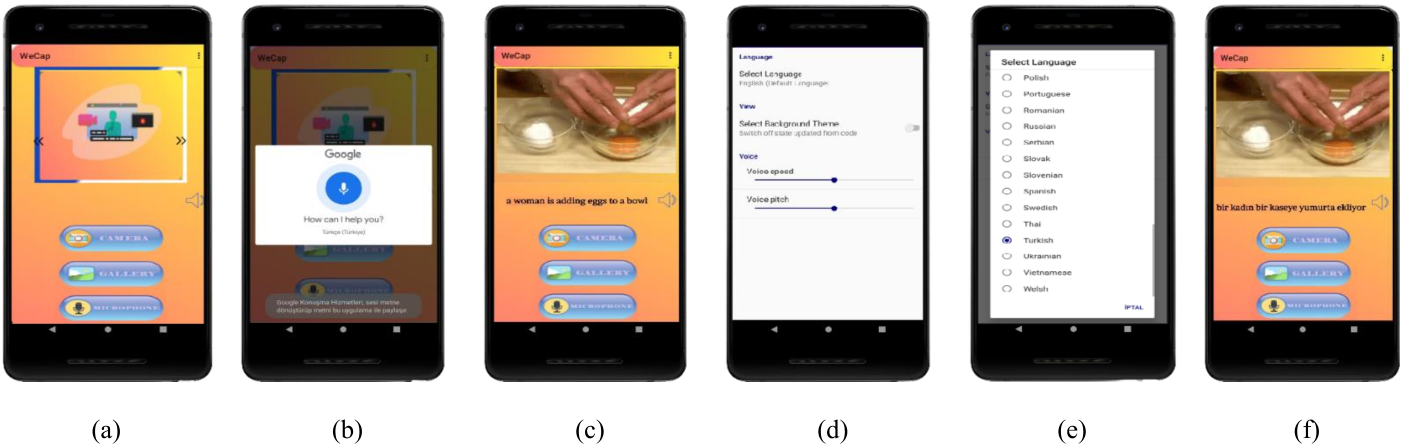


Figure 2. Android application: the homepage is given in (a), voice command is shown in (b), the generated caption with video is shown in (c), settings, language options and the translated caption are given in (d), (e) and (f).

The performance metrics BLEU-n (n = 1, 2, 3, 4), METEOR, and ROUGE-L, evaluate the similarity between a machine-generated caption and the reference captions in the machine translation systems. SPICE is designed to evaluate captioning tasks that compare the semantic propositional content of generated and reference captions. CIDEr is also designed for captioning, examining the average cosine similarity between the generated and the reference captions. The results are sorted based on the CIDEr metric due to its better correlation with human judgment than BLEU-n, METEOR, SPICE, and ROUGE-L.

3.2. Results and Discussion

In this study, an encoder-decoder based sequence-to-sequence system based on multi-layer GRU has been presented for video captioning. The ResNet-101 with multi-layer GRU was evaluated on MSVD in terms of BLEU-n, CIDEr, METEOR, ROUGE-L, and SPICE. The experimental results in

Table 1 examine the effect of layer number for multi-layer GRU architecture. Regarding the six different configurations for the number of layers, 2-layer GRU architecture generally outperforms the other schemes. For BLEU-n and CIDEr metrics, 2-layer GRU architecture yields the highest performance, while 3-layer GRU architecture yields higher results in terms of CIDEr and SPICE metrics. The empirical results indicate that the performance of the captioning systems degrades after 4-layer GRU architecture. Based on the empirical results among all the compared configurations, 2-layer GRU with ResNet-101 has been integrated into the *WeCap* Android application. In Table 1, the proposed video captioning system has been compared with five state-of-the-art architectures on the MSVD dataset in terms of six evaluation metrics. The proposed video captioning architecture which combines ResNet-101 with 2-layer GRU outperforms the compared state-of-the-art architectures in terms of five evaluation metrics, i.e., the highest CIDEr, BLEU-4, BLEU-3, BLEU-2, and BLEU-1 scores have been achieved by this architecture. The second highest METEOR score among all the compared schemes has been obtained by the proposed architecture. In Table 3, examples of ground truth and generated captions by the proposed system for two videos have been presented. The empirical results listed in Table 2 and the captions presented in Table 3 indicate that the proposed architecture can yield promising results and can be utilized as a viable tool for video captioning.

4. Conclusions

In this paper, a video captioning system has been developed using an encoder-decoder based sequence-to-sequence approach. ResNet-101 CNN architecture was 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 on the MSVD dataset indicate that the proposed approach can generate meaningful captions with multi-layer GRU. Then, the proposed system is integrated with our custom-designed Android application, called *WeCap*, to bring it into practical use. Besides, *WeCap* extracts features and generates captions via the embedded encoder-decoder in offline mode. The empirical results indicate that the proposed architecture is capable of producing promising results and is a viable tool for video captioning. For further research, the video captioning approach will be optimized with hyper-parameters and the addition of residual connections.

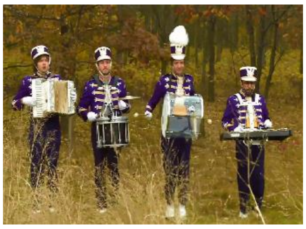
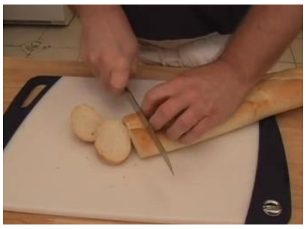
Table 1. Comparison of Different Layer-Sized GRU with ResNet-101

Design	# of Layers	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	SPICE
ResNet-101 with Multi-layer GRU	1	0.646	0.482	0.602	0.704	0.823	**0.698**	**0.330**	0.048
	2	**0.698**	**0.516**	**0.621**	**0.718**	**0.831**	0.695	0.329	0.047
	3	**0.698**	0.489	0.590	0.689	0.815	0.692	0.319	**0.048**
	4	0.608	0.433	0.537	0.645	0.780	0.675	0.310	0.042
	5	0.601	0.423	0.525	0.639	0.765	0.658	0.301	0.042
	6	0.594	0.414	0.515	0.620	0.754	0.645	0.299	0.040

Table 2. Comparison of our proposed system with state-of-the-art architectures on MSVD dataset

	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	METEOR
(P. Pan et al., 2016)	-	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, Wang, Huang, Yang, & Xu, 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 ResNet-101 with 2-layer GRU	**0.698**	**0.516**	**0.621**	**0.718**	**0.831**	0.329

Table 3. Examples of ground truth and generated captions of video frames selected from the MSVD dataset

	
Reference Captions: **(1)** A band of four men are playing music in an outdoor location. **(2)** A band of four people are playing different instruments. **(3)** A four person band is playing. **(4)** Four men are playing outdoors in a marching band. **(5)** Four people playing instruments are marching in place.	**(1)** A person with a knife is slicing bread. **(2)** A man is slicing some bread. **(3)** The man sliced the loaf of bread. **(4)** Man cutting a piece of bread with knife. **(5)** Someone is slicing pieces of bread from a loaf of french bread.
Generated Captions: **1-layer:** two men are playing **2-layer:** a band is performing on stage **3-layer:** a man is playing a **4-layer:** a man is dancing **5-layer:** a man is playing a horse **6-layer:** a man is riding a horse	**1-layer:** a man is slicing bread **2-layer:** a man is slicing a piece of bread **3-layer:** a man is slicing a carrot **4-layer:** a woman is cutting an apple **5-layer:** a man is cutting an onion **6-layer:** a person is cutting an onion

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).

References

Amaresh, M., & Chitrakala, S. (2019). *Video captioning using deep learning: An overview of methods, datasets and metrics.* Paper presented at the 2019 International Conference on Communication and Signal Processing.

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

Banerjee, S., & Lavie, A. (2005). *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* Paper presented at the Proceedings of the 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.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Çaylı, Ö., Makav, B., Kılıç, V., & Onan, A. (2020). *Mobile Application Based Automatic Caption Generation for Visually Impaired.* Paper presented at the International Conference on Intelligent and Fuzzy Systems.

Chen, D., & Dolan, W. B. (2011). *Collecting highly parallel data for paraphrase evaluation.* Paper presented at the 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.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

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.* Paper presented at the 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.* Paper presented at the 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.* Paper presented at the Proceedings of the 24th ACM International Conference on Multimedia.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Keskin, R., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). *Multi-GRU Based Automated Image Captioning for Smartphones.* Paper presented at the 2021 29th Signal Processing and Communications Applications Conference

Kılıç, V. (2021). Deep Gated Recurrent Unit for Smartphone-Based Image Captioning. *Sakarya University Journal of Computer Information Sciences, 4*(2), 181-191.

Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries.* Paper presented at the Text summarization branches out.

Makav, B., & Kılıç, V. (2019). *Smartphone-based image captioning for visually and hearing impaired.* Paper presented at the 11th International Conference on Electrical and Electronics Engineering

Pan, P., Xu, Z., Yang, Y., Wu, F., & Zhuang, Y. (2016). *Hierarchical recurrent neural encoder for video representation with application to captioning.* Paper presented at the Proceedings of the IEEE 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.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). *A dataset for movie description.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Shen, F., Shen, C., Shi, Q., Van Den Hengel, A., & Tang, Z. (2013). *Inductive hashing on manifolds.* Paper presented at the 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.* Paper presented at the 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:.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:.01070*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). *Cider: Consensus-based image description evaluation.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). *Sequence to sequence-video to text.* Paper presented at the 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.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Xu, R., Xiong, C., Chen, W., & Corso, J. (2015). *Jointly modeling deep video and compositional text to bridge vision and language in a unified framework.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). *Describing videos by exploiting temporal structure.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). *Video paragraph captioning using hierarchical recurrent neural networks.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.