Avrupa Bilim ve Teknoloji Dergisi
Sayı 35, S. 380-386, Nisan 2022
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 35, pp. 380-386, April 2022
Copyright © 2022 EJOSAT
Research Article

Sequence-to-Sequence Video Captioning with Residual Connected Gated Recurrent Units

Selman Aydın[1*], Özkan Çaylı[1], Volkan Kılıç[1], Aytuğ Onan[2]

[1] Izmir Katip Celebi University, Faculty of Engineering, Departmant of Electrical and Electronics, İzmir, Turkey, (ORCID: 0000-0002-2851-6303, 0000-0002-3389-3867, 0000-0002-3164-1981), selman.aydin017@gmail.com, ozkan.cayli@ikcu.edu.tr, volkan.kilic@ikcu.edu.tr

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

(First received 11 February 2022 and in final form 16 April 2022)

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

ATIF/REFERENCE: Aydın, S., Çaylı, Ö., Kılıç, V., & Onan, A. (2022). Sequence-to-Sequence Video Captioning with Residual Connected Gated Recurrent Units. *European Journal of Science and Technology*, (35), 380-386.

Abstract

Recurrent neural networks, have recently emerged as a useful tool in computer vision and language modeling tasks such as image and video captioning. The main limitation of these networks is preserving the gradient flow as the network gets deeper. We propose a residual connection based video captioning approach to overcome this limitation and maintain the gradient flow by carrying the information through layers from bottom to top with additive features. The experimental evaluations on the MSVD dataset indicate that the proposed approach achieves accurate caption generation compared to the state-of-the-art results. In addition, the proposed approach is integrated with our custom-designed Android application, *WeCapV2*, capable of generating captions without an internet connection.

Keywords: Convolutional Neural Network, Recurrent Neural Network, Residual Connections, Video Captioning, Android Application

Artık Bağlı Kapılı Tekrarlayan Birimlerle Sıradan Sıraya Video Altyazılama

Öz

Tekrarlayan sinir ağları, son zamanlarda görüntü ve video altyazılama gibi bilgisayarla görme ve dil modelleme görevlerinde kullanışlı bir araç olarak ortaya çıkmıştır. Bu ağların ana sorunu ağ derinleştikçe gradyan akışını koruyamamaktır. Bu sorunun üstesinden gelmek ve gradyan akışını sürdürmek için bilgileri katmanlar arasında aşağıdan yukarıya özellikleri taşıyan artık bağlantı tabanlı bir video altyazı yaklaşımı öneriyoruz. MSVD veriseti üzerindeki deneysel değerlendirmeler, önerilen yaklaşımın en son sonuçlarla karşılaştırıldığında doğru altyazı oluşturmayı başardığını göstermektedir. Ayrıca, önerilen yaklaşım, özel tasarım olan Android uygulamamız *WeCapV2* ile entegre edilip internet bağlantısı olmadan altyazı oluşturabilmektedir.

Anahtar Kelimeler: Evrişimsel Sinir Ağı, Kapılı Tekrarlayan Birim, Artık Bağlantılar, Video Altyazılama, Android Uygulama

[*] Corresponding Author: selman.aydin017@gmail.com

1. Introduction

Video captioning aims to describe video content with a meaningful caption using techniques from the fields of natural language processing and computer vision. There has been increasing attention for video captioning due to its potential in video understanding (Gan, Yao, Yang, Yang, & Mei, 2016) and virtual assistant applications (Baran, Moral, & Kılıç, 2021; Çaylı, Makav, Kılıç, & Onan, 2020; Fetiler, Çaylı, Moral, Kılıç, & Onan, 2021; Keskin, Çaylı, Moral, Kılıç, & Onan, 2021; Makav & Kılıç, 2019a, 2019b).

Earlier studies mostly employed the template-based approach, which uses fixed templates that are a set of most likely words to generate a caption (Guadarrama et al., 2013; Khan, Zhang, & Gotoh, 2011; Rohrbach et al., 2013). In other words, a sentence is first parsed into an SVO triplet that stands for the subject, verb, and object, respectively. Then using continuous space word representation (Frome et al., 2013), each word in SVO is expressed as a continuous vector. The norm of continuous vectors denotes the correlation between words and continuous vector form of SVO defines sentence templates. Then, various objects, attributes and actions are detected from frames to generate a syntactically correct caption with predefined sentence templates. Therefore, the performance of those methods highly depends on a predefined template and are limited by the accuracy of the detection of the word.

Recently, a deep learning-based encoder-decoder framework was proposed for more accurate captions than the template-based approach (Amirian, Rasheed, Taha, & Arabnia, 2020). This framework combines a convolutional neural network (CNN) to extract features from each frame of an input video in the encoder and a recurrent neural network (RNN) in the decoder to generate caption. There are several pre-trained CNNs, including ResNet (Targ, Almeida, & Lyman, 2016), Xception (Chollet, 2017), and Inception-v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), which are commonly used in the encoder. The RNN-based decoder processes extracted features by the encoder to generate caption word-by-word. However, the simple RNN suffers from vanishing and exploding gradient problems, causing short-term dependencies.

Gating mechanisms in RNN are introduced to mitigate this issue, such as long short-term memory (LSTM) and gated recurrent unit (GRU). LSTM has three gates to regulate the flow of information and cell state to keep information in memory for long periods. Similar to the LSTM unit, the GRU has two gates to control the flow of information, but no cell state.

In (Venugopalan et al., 2014), the sequence-to-sequence approach was proposed that uses CNN to extract features in the encoder, and LSTM for generating natural language captions in the decoder. Therefore, the temporal structures of the video sequence were ignored in this approach. Yao et al. introduced a temporal attention mechanism based on a soft-alignment method to evaluate the importance of each frame (L. Yao et al., 2015). Donahue et al. used an LSTM architecture to model visual time series, and used the maximum posterior estimation of a conditional random field to define the semantic representation of the video in the tuple (Donahue et al., 2015). Finally, it converted the tuple into sentences with the final LSTM layer. (Venugopalan et al., 2015) proposed a completely neural architecture approach that includes temporal information with the optical flow in the encoder and stacked LSTMs in both encoder and decoder. Before passing through a CNN, optical flow between pairs of frames was calculated to model the temporal aspects of activities. Stacked LSTM encoded the CNN outputs of each frame, and LSTM in the decoder was conditioned on the last hidden state of the LSTM in the encoder to generate caption. It is the first implementation of sequence to sequence approach on video captioning as opposed to conventional implementation in machine translation (T. Yao, Pan, Li, Qiu, & Mei, 2017).

Although the deep learning-based encoder-decoder framework has been found to be promising, it limits preserving the gradient flow as the network gets deeper. In (He, Zhang, Ren, & Sun, 2016), a deep residual learning framework is presented to solve the vanishing gradient problem in depth networks. Residual nets, which are eight times deeper than VGG nets, is evaluated on the ImageNet dataset, and improved accuracy is achieved for image classification and object detection solely. In (Wu et al., 2016), an end-to-end learning approach is proposed for automated translation that residual connected LSTM network used on both encoder and decoder.

In this study, a residual connection based sequence-to-sequence video captioning approach is proposed to mitigate the vanishing gradient problem in deep layers. The encoder of the proposed approach consists of a combination of the Inception-v3 CNN architecture to extract features from the frames one by one and a residual connected multi-layer GRU to encode features, while the decoder employs multi-layer GRU with residual connections to generate meaningful caption.

The proposed approach was trained on the MSVD dataset and evaluated with commonly used performance metrics, such as BLEU-n (n = 1, 2, 3, 4) (Papineni, Roukos, Ward, & Zhu, 2002), ROUGE-L (Lin, 2004), SPICE (Anderson, Fernando, Johnson, & Gould, 2016), METEOR (Banerjee & Lavie, 2005) and CIDEr (Vedantam, Lawrence Zitnick, & Parikh, 2015). The impact of residual connections on captioning performance is analyzed with these metrics. Then, 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 covers the encoder-decoder-based sequence-to-sequence approach for video captioning. Section 3 presents the dataset, performance metrics, and proposed approach results. Closing remarks are given in Section 4.

2. Methodology

In this section, the proposed video captioning approach based on sequence-to-sequence learning with residual connections is presented. Then, our custom-designed Android application, *WeCapV2*, capable of running the proposed approach in offline mode (without internet connection), is described.

2.1. Proposed Video Captioning Approach

The proposed video captioning approach is based on an encoder-decoder framework, in which the encoder extracts visual attributes from each frame to feed the decoder that creates captions describing events and scenes for relevant parts of the video. As illustrated in Figure 1, the encoder contains both CNN and RNN while the decoder consists an RNN, embedding and fully-connected layers. Both encoder and decoder use multi-layer

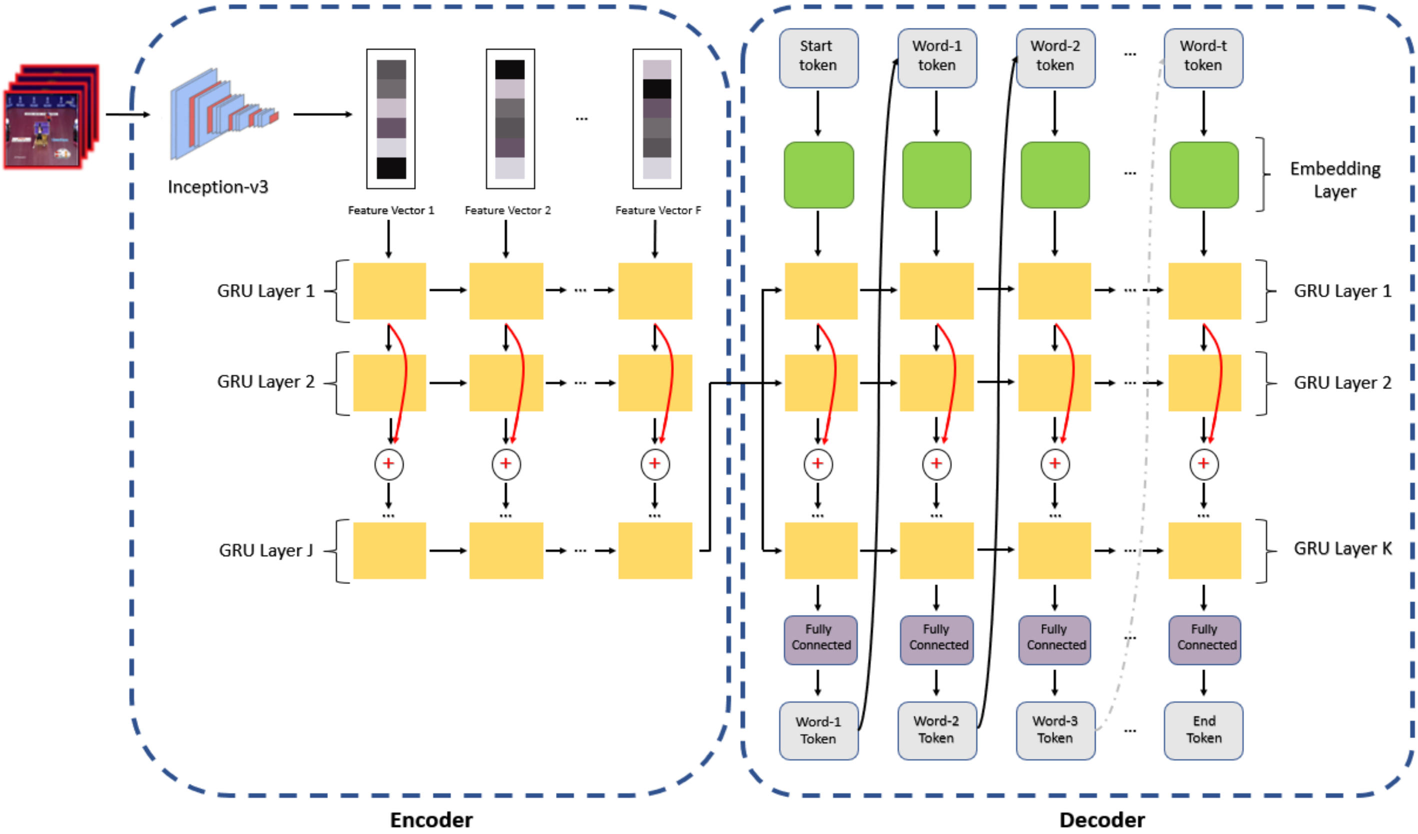


Figure 1. The proposed Encoder-Decoder Approach

GRU as an RNN due to its ability to retain more semantic information of the image (Keskin, Moral, Kılıç, & Onan, 2021; Kılıç, 2021). The proposed encoder first utilizes a pre-trained CNN architecture to extract attributes. The purpose of the multi-layer GRU based RNN in the encoder is to create a final feature vector for the decoder. Then, the multi-layer GRU based RNN in the decoder processes the feature vector with embedding and fully-connected layers to generate a caption.

GRU is an improved version of RNN that solves the vanishing and exploding gradient problem for long sequence data. It has a gating mechanism consisting of an update (z_t) and reset gates (r_t), which control the flow of information through cells (W. Liu, Wang, Zhu, & Chen, 2020). The update gate (z_t) decides which information should be passed to the next step while the reset gate (r_t) controls what information should be forgotten. The flow of information in GRU has been carried on with the following equations at time step t:

$$r_t = \sigma(W_r x_t + W_r h_{t-1}) \qquad (1)$$

$$z_t = \sigma(W_z x_t + W_z h_{t-1}) \qquad (2)$$

$$u_t = tanh(W_h x_t + W_h(r_t \odot h_{t-1})) \qquad (3)$$

$$h_t = (1 - z_t)h_{t-1} + z_t u_t \qquad (4)$$

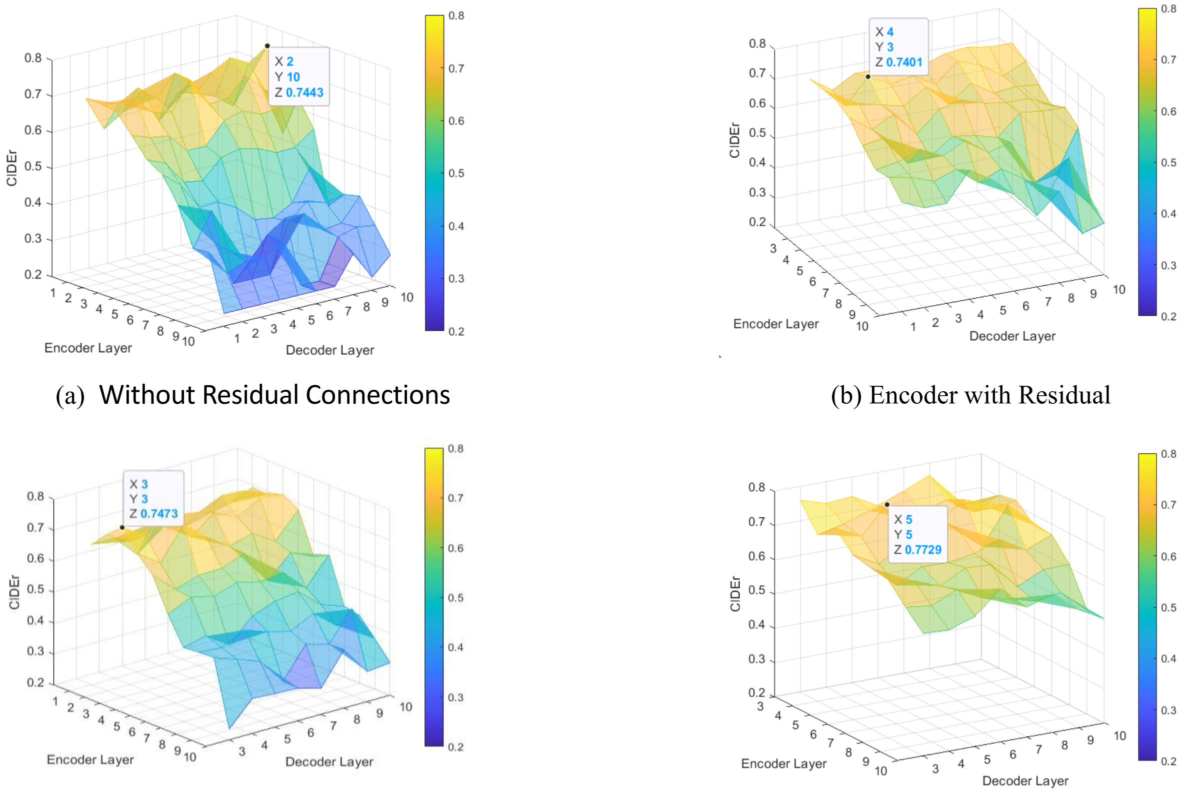
where x_t is the input vector, h_{t-1} is the hidden state vector of the previous time step, u_t is the candidate hidden vector and W refers to the weights. σ and *tanh* are sigmoid and hyperbolic tangent activation functions. $\odot$ denotes the Hadamard (element-wise) product operator. The residual connections allow us to train very deep encoder and decoder networks as they significantly improve the gradient flow in the backward pass (Wu et al., 2016). Residual connections are shown with red arrows in Figure 1 and the flow of the input on GRU is as follows:

$$x_t^i = x_t^{i-1} \oplus x_t^{i-2} \qquad (5)$$

where x_t^i is the input vector of GRU at time step t, and i represents the respective layer of GRU, starting from 3. The input of (i)-th GRU layer is the element-wise sum of the input of ($i-1$)-th and ($i-2$)-th GRU layers. The proposed encoder employs the Inception-v3 CNN architecture, which consists of 48 convolution, pooling, and fully connected layers, for feature extraction in video frames. Multi-layer GRU uses these feature vectors as input to represent them with a single vector. The last hidden state in the last GRU layer is fed to the decoder for further process of caption generation. The multi-layer GRU in the decoder uses these vectors as input. The final output of the last GRU layer is fed into the fully connected layer. The FC layer generates the word-1 token used by the embedding layer in the next time step, and this process is repeated t times until it reaches the end token. A fully connected layer predicts meaningful and grammatically correct words at each time step to generate a caption.

2.1. Android Application

The proposed approach was embedded in our custom-designed Android application, *WeCapV2*, capable of generating captions in offline mode. First, the proposed video captioning approach was optimized using PyTorch (Paszke et al., 2019), an open-source machine learning framework to perform inference on mobile devices. Dynamic quantization method was applied to reduce the size of model weight, leading to a faster execution time



(a) Without Residual Connections

(b) Encoder with Residual

(c) Decoder with Residual

(d) Encoder-Decoder with Residual

Figure 2. 3D illustration of CIDEr results. The residual connections are not applied in (a) while the residual connections are used on the encoder in (b). The decoder has residual connections in (c), and the residual connections are applied on both encoder and decoder in (d).

for video captioning. To integrate the quantized model into the application, it was converted to TorchScript format. Two buttons on the homepage allow the user to choose a video from the gallery or capture a video. An icon on the button indicates these preferences. After the user uploads the video, the embedded model generates captions directly without requiring an internet connection. The generated caption with running time is displayed on the homepage under the video. Although the caption is generated in English as a default, the application allows translating other languages using the language settings of the smartphone. After the language selection, the caption could be translated using Google Cloud Translation API.

3. Experimental Evaluations

The dataset and performance metrics, implementation details, and evaluation of the proposed video captioning approach are presented in this section.

3.1. Dataset and Performance Metrics

In order to evaluate the proposed approach, three datasets, including a large set of videos and reference captions, have been analysed. Youtube2Text (Chen & Dolan, 2011), MSR-VTT (Xu, Mei, Yao, & Rui, 2016), and MSVD (Chen & Dolan, 2011) are open source video captioning datasets. Youtube2Text includes 1970 video clips from YouTube consisting of 1300 training, 1300 validation and 670 test with a total number of 80839 reference captions. MSR-VTT contains 6513 training, 2990 validation and 497 test video clips with an average of 20 captions for each video.

Datasets	Train	Validation	Test	Total Caption
Yotube2Text	1300	1300	670	80839
MSR-VTT	6513	2990	497	20000
MSVD	1380	295	295	78800

Table 1 Comparison of Datasets

MSVD consists of 1970 short video clips collected from YouTube while there are multi-language descriptions for videos, and there is an average of 40 English descriptions for a video clip. In this study, MSVD is adopted to evaluate our video captioning approach due to its large reference caption set compared to other datasets. Furthermore, we split the MSVD into 1380 training, 295 validation and 295 test video clips. The specifications of the datasets are shown in Table 1.

The study is evaluated on common performance metrics such as BLEU-n, METEOR, ROUGE-L, SPICE, and CIDEr. BLEU-n and METEOR are used to evaluate machine translation. BLEU-n uses n-gram to evaluate generated caption with the reference caption, while METEOR aligns translation hypotheses with reference translations and evaluates them by calculating sentence-level similarity scores. ROUGE-L compares reference and generated summary based on the longest subsequence. SPICE is a semantic evaluation metric and evaluates the objects, attributes, and relationships in the generated captions, rather than comparing the generated captions with reference captions for syntactic agreement (S. Liu, Zhu, Ye, Guadarrama, & Murphy, 2017). CIDEr evaluates the consensus between the generated caption and

Table 2 Performance Metric Results

	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	SPICE
Encoder with residual connections	0.740	0.502	0.601	0.697	0.828	0.700	0.331	0.048
(Pan, Yao, Li, & Mei, 2017)	0.740	0.528	**0.628**	**0.720**	0.828	-	**0.335**	-
Without residual connections	0.744	**0.536**	**0.628**	0.715	**0.855**	0.704	0.327	0.051
Decoder with residual connections	0.747	0.482	0.589	0.687	0.812	0.698	0.328	0.050
(Gao, Guo, Zhang, Xu, & Shen, 2017)	0.748	0.508	0.611	0.708	0.818	-	0.333	-
Encoder-Decoder residual connections	**0.772**	0.502	0.602	0.702	0.823	**0.706**	0.330	**0.053**

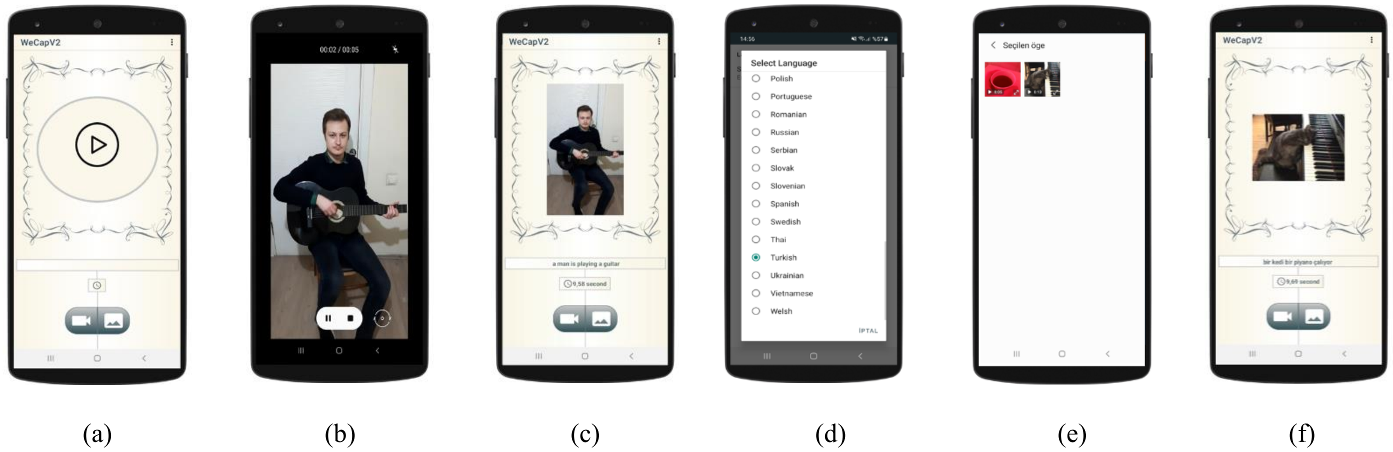


Figure 3. Android application: the homepage is given in (a), a live video is captured in (b), the generated caption with video is shown in (c), language options, gallery, and the translated caption are given in (d), (e) and (f), respectively.

reference captions. Results in this paper are ranked by the CIDEr metric as it better evaluates salient features in the image, and relationship between the generated caption and reference captions grammatically and semantically.

3.2. Implementation Details

We split the input video into eight equally distributed parts to take a sample frame from each part. Therefore, the number of feature vectors denoted as F in Figure 1 becomes eight. Latent feature size in the embedding, fully connected, and GRU layers were set to 256. After processing and tokenizing the captions in the training set, the vocabulary size was set to 1270, resulting in a same size for the fully connected layer. Last, a stochastic gradient descent optimizer was employed in training with a 0.01 learning rate.

3.3. Result and Discussion

Four different encoder-decoder designs depending on the residual connections have been examined to observe the effect of residual connections in deep layers. All designs were evaluated with BLEU-n, ROUGE-L, SPICE, METEOR, and CIDEr metrics on the MSVD dataset. CIDEr performance of different encoder-decoder designs has been visualized with mesh graph in Figure 2. The encoder and decoder layer numbers are specified on horizontal axes, and CIDEr metric results are specified on vertical axes. The visualized CIDEr results in Figure 2 indicate that the performance of the captioning system degrades when the layer gets deeper without residual connections, and using the residual connections in both encoder and decoder prevents this issue and leads to an improved performance compared to the others. Among all the designs, the highest performance is obtained by 5-layer GRU using the residual connections in both encoder and decoder. Therefore, it has been integrated into the *WeCapV2* application.

In Table 2, the proposed approach is also compared with state-of-the-art approaches on the MSVD dataset in terms of six evaluation metrics. The results in Table 2 indicate that the proposed approach, in which the residual connections are used in both encoder and decoder, outperforms the state-of-the-art approaches in terms of the CIDEr metric. Table 3 shows the ground truth and generated captions by the proposed approach for two videos. Note that generated caption by our proposed approach is more meaningful compared to other encoder-decoder designs.

Screen shots of the application are given in Figure 3. The homepage welcomes the user at the first login in Figure 3 (a). Next, the user accesses the video capture screen by tapping the camera icon in Figure 3 (b). Video capture time is limited to a maximum of 5 seconds. Figure 3 (c) shows the generated caption from the captured video. The language selection for the caption is shown in Figure 3 (d). Gallery icon is used to choose a video from the gallery in Figure 3 (e). Figure 3 (f) shows the generated caption for the video taken from the gallery.

4. Conclusions

In this paper, an encoder-decoder-based sequence-to-sequence video captioning approach has been presented. This approach is based on Inception-v3 CNN to extract features from video frames, and residual connected multi-layer GRU was used to process the features and generate the caption. The results on the MSVD dataset demonstrate that the proposed approach can achieve meaningful and grammatically true caption with residual

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

Reference Captions: **(1)** A soccer team practicing drills. **(2)** A few men are playing football on a field. **(3)** A soccer team kicks a soccer ball around. **(4)** Men are playing soccer. **(5)** Men are practicing soccer.	**Reference Captions:** **(1)** Three women are dancing outside. **(2)** The ladies danced in white dresses outside. **(3)** The three girls danced on the grass. **(4)** Three women are dancing in a field. **(5)** Three women dance in a green field.
Generated Captions: **No Residual:** A group of the playing the ball. **Encoder Residual:** The boys are playing the ball. **Decoder Residual:** A group is playing the ball. **Encoder-Decoder Residual:** A group of men are playing soccer.	**No Residual:** A man is dancing. **Encoder Residual:** A group of men are fighting. **Decoder Residual:** A group of men are dancing. **Encoder-Decoder Residual:** A group of people are dancing.

connections. Our proposed approach achieves state-of-the-art performance in terms of CIDEr. Additionally, we integrated the approach into our custom-designed Android application, *WeCapV2*, which extracts features and generates captions with an embedded encoder and decoder. Future research includes bidirectionality in recurrent neural networks and attention mechanisms to increase the accuracy of captions in terms of performance metrics.

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 TUBITAK 2209-B Industry Oriented Research Project Support Programme for Undergraduate Students with project no: 1139B412100439.

References

Amirian, S., Rasheed, K., Taha, T. R., & Arabnia, H. R. J. I. A. (2020). Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. *IEEE Access, 8*, 218386-218400.

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.

Baran, M., Moral, Ö. T., & Kılıç, V. J. A. B. v. T. D. (2021). Akıllı telefonlar için birleştirme modeli tabanlı görüntü altyazılama. *European Journal of Science and Technology*(26), 191-196.

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.

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.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Fetiler, B., Çaylı, Ö., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). Video captioning based on multi-layer gated recurrent unit for smartphones. *European Journal of Science and Technology*(32), 221-226.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. J. A. i. n. i. p. s. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems, 26*.

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.

Gao, L., Guo, Z., Zhang, H., Xu, X., & Shen, H. T. J. I. T. o. M. (2017). Video captioning with attention-based LSTM and semantic consistency. *IEEE Transactions on Multimedia, 19*(9), 2045-2055.

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.

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., Çaylı, Ö., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). A benchmark for feature-injection architectures in image captioning. *European Journal of Science and Technology*(31), 461-468.

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.

Khan, M. U. G., Zhang, L., & Gotoh, Y. (2011). *Human focused video description.* Paper presented at the 2011 IEEE International Conference on Computer Vision Workshops.

Kılıç, V. (2021). Deep gated recurrent unit for smartphone-based image captioning. *Sakarya University Journal of Computer and 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.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). *Improved image captioning via policy gradient optimization of spider.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Liu, W., Wang, Q., Zhu, Y., & Chen, H. J. T. J. o. S. (2020). GRU: optimization of NPI performance. *The Journal of Supercomputing, 76*(5), 3542-3554.

Makav, B., & Kılıç, V. (2019a). *A new image captioning approach for visually impaired people.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering.

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

Pan, Y., Yao, T., Li, H., & Mei, T. (2017). *Video captioning with transferred semantic attributes.* 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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Antiga, L. J. A. i. n. i. p. s. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013). *Translating video content to natural language descriptions.* Paper presented at the Proceedings of the IEEE international Conference on Computer Vision.

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. J. a. p. a. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

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. J. a. p. a. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . Macherey, K. J. a. p. a. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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.

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.

Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). *Boosting image captioning with attributes.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.