Knowledge Distillation for Efficient Audio-Visual Video Captioning

Özkan Çaylı[1], Xubo Liu[2], Volkan Kılıç[1*], Wenwu Wang[2†]
[1]Electrical and Electronics Engineering, İzmir Katip Çelebi University, Türkiye
[2]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
Email: *volkan.kilic@ikcu.edu.tr; †w.wang@surrey.ac.uk

***Abstract*—Automatically describing audio-visual content with texts, namely video captioning, has received significant attention due to its potential applications across diverse fields. Deep neural networks are the dominant methods, offering state-of-the-art performance. However, these methods are often undeployable in low-power devices like smartphones due to the large size of the model parameters. In this paper, we propose to exploit simple pooling front-end and down-sampling algorithms with knowledge distillation for audio and visual attributes using a reduced number of audio-visual frames. With the help of knowledge distillation from the teacher model, our proposed method greatly reduces the redundant information in audio-visual streams without losing critical contexts for caption generation. Extensive experimental evaluations on the MSR-VTT dataset demonstrate that our proposed approach significantly reduces the inference time by about 80% with a small sacrifice (less than 0.02%) in captioning accuracy.**

***Index Terms*—Image Processing, Audio Processing, Natural Language Processing, Deep Learning, Video Captioning**

I. Introduction

Video captioning aims to generate grammatically and semantically meaningful sentences for the content of audio-visual media, driven by applications such as video indexing or retrieval and virtual assistants for visually and hearing-impaired people [1], [2].

This task involves several challenges, such as identifying objects and scenes in the video frame, extracting audio attributes, and audio-visual fusion to describe the content with certain grammatical structures and semantics [3]–[6]. These issues could be addressed with the release of large-scale datasets and advances in deep learning, which has led to the development of highly complex networks with improved caption generation. However, this can also lead to high computational cost due to the increased complexity of the networks and scale of the datasets. One approach to overcome this issue is to use efficient audio and visual feature extraction networks as they provide faster inference time [7]. These networks can be categorized into four classes: namely, model compression [8], [9], knowledge distillation [10]–[12], efficient networks [13], [14], and simple pooling front-ends (SimPFs) [15]. A framework that applies passive filter pruning to reduce the number of convolutional filters is proposed for a compressed convolutional neural network (CNN) [8]. Similarly, a low-complexity CNN architecture is presented in [9], by reducing model parameters and memory usage. A BERT architecture is proposed as a teacher network that provides soft labels to guide a seq2seq network for audio speech recognition [10]. In a highway deep neural network, knowledge distillation and teacher-student training are leveraged to achieve improved accuracy with a reduced number of parameters [11]. Pretrained audio neural networks (PANNs) [13], which are trained on AudioSet [16], can be transferred to audio-related tasks such as audio classification and captioning [17]–[19]. SimPFs are employed to reduce the required number of audio frames by reducing floating point operations on a network for efficient audio classification [15].

For visual feature extraction, knowledge distillation is used in [20] to generate soft labels for simpler networks to be deployed on a device with low computing resources. Similarly, knowledge distillation with an attention mechanism is used in [21], which groups high-dimensional features into low-dimensional vectors. Furthermore, [22] uses all the visual frames in a video to train the teacher network. The student network then uses uniformly down-sampled frames and mimics the teacher for efficient video classification.

In this study, we propose an efficient audio-visual captioning method based on the teacher-student network, which uses knowledge distillation for audio and visual feature extraction with a reduced number of frames, leading to substantially improved captioning efficiency. More specifically, the PANNs network [13] is used with SimPF [15] for audio feature extraction, while Inception-v3 CNN architecture [23] with down-sampling is utilized for visual feature extraction [24]. The language model uses simple stacked gated recurrent units (GRUs) [25] with dropouts [26] and residual connections [27], [28]. The student network is first trained and fine-tuned with the cross-entropy loss. To further improve the captioning accuracy, the representation loss is also used along with the cross-entropy loss. The experiments show that knowledge distillation can speed up audio-visual feature extraction with a negligible drop in captioning accuracy.

This paper is organized as follows: Section II presents the proposed audio-visual video captioning approach. Section III describes the dataset and performance metrics for experimental evaluations and discusses the experimental results, followed by the conclusions.

II. Proposed Approach

This section presents the proposed video captioning approach based on the teacher-student model, as illustrated in

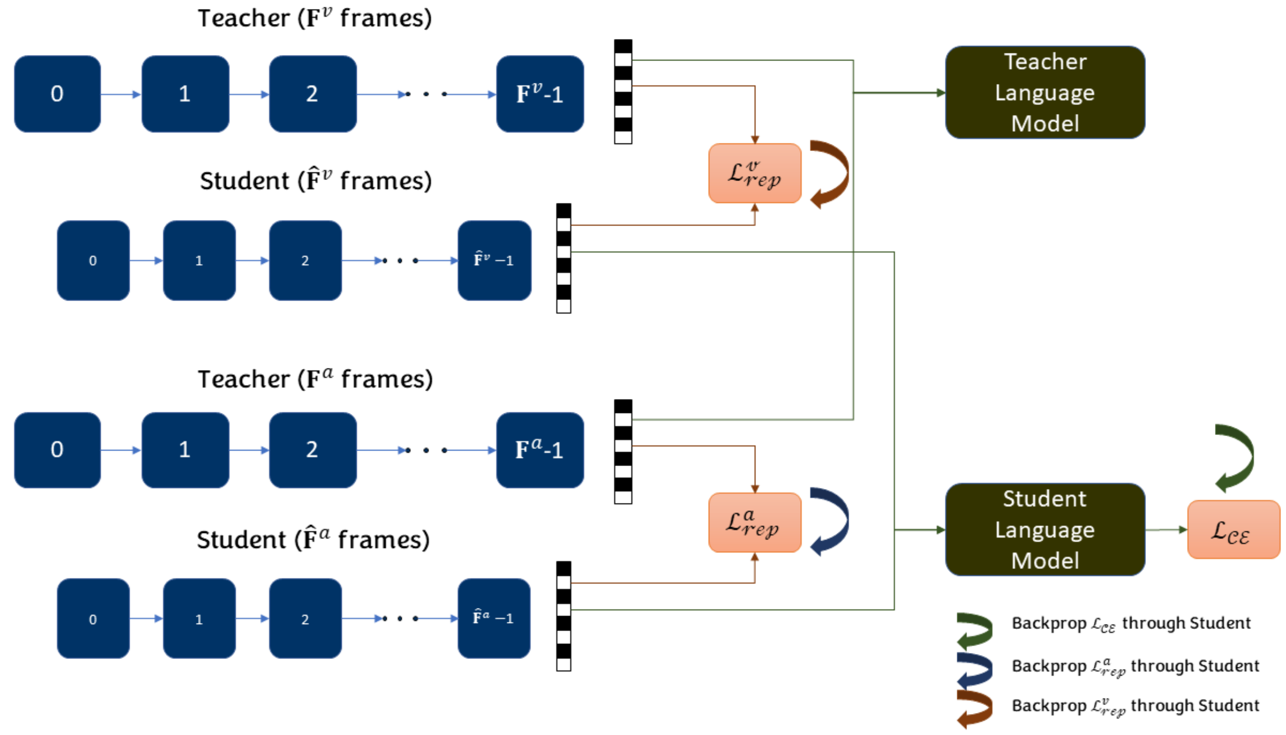
Teacher ($\mathbf{F}^v$ frames)
0
1
2
$\mathbf{F}^v$-1
Student ($\hat{\mathbf{F}}^v$ frames)
0
1
2
$\hat{\mathbf{F}}^v$ −1
Teacher ($\mathbf{F}^a$ frames)
0
1
2
$\mathbf{F}^a$-1
Student ($\hat{\mathbf{F}}^a$ frames)
0
1
2
$\hat{\mathbf{F}}^a$ −1
$\mathcal{L}^v_{rep}$
$\mathcal{L}^a_{rep}$
Teacher Language Model
Student Language Model
$\mathcal{L}_{CE}$
Backprop $\mathcal{L}_{CE}$ through Student
Backprop $\mathcal{L}^a_{rep}$ through Student
Backprop $\mathcal{L}^v_{rep}$ through Student

Fig. 1: The Proposed Approach

Figure 1.

In video captioning, a sequence of words needs to be predicted from a vocabulary using audio and visual attributes. The teacher network utilizes N^a audio frames $\mathbf{F}^a = (F^a_0, F^a_1, ...F^a_{N_a-1})$ and N^v video frames $\mathbf{F}^v = (F^v_0, F^v_1, ...F^v_{N_v-1})$ of the video $\mathbf{V}$ to predict a caption which can be stated using a neural network f:

$$P(\hat{\mathbf{Y}}|\mathbf{V}) = f(\mathbf{F}^a, \mathbf{F}^v), \tag{1}$$

where $\hat{\mathbf{Y}}$ denotes a series of words as $(\hat{y}_0, \hat{y}_1, ...\hat{y}_{N^c})$ and N^c refers to the number of words in the caption.

We employ Inception-v3 CNN architecture pre-trained on the ImageNet dataset to extract features from visual frames. The architecture resizes the images to 3×299×299, then the average pooling layer outputs a latent vector consisting of 2048 units. Similarly, audio features are extracted with PANNs CNN architecture containing 10 stacked CNN layers pre-trained on AudioSet. A recurrent neural network (RNN)-based network that utilizes audio and visual features from the Inception-v3 and PANNs is used as a language model to generate captions. We employ a mean operator and acquire latent vectors from time-series input, which describe audio and visual features. These latent vectors are concatenated and fed to the RNN-based network consisting of embedding, GRUs, and linear layers. Moreover, residual connections and dropouts are applied between layers to maintain gradient flow from the lower to upper layers. The teacher network is trained with the cross-entropy loss denoted as $\mathcal{L}_{CE}$. The student network is similar to the teacher, where SimPF and down-sampling algorithms are employed to reduce the number of audio and visual frames by a compression rate in a video. Specifically, we use the spectral pooling method of SimPF, which computes the discrete Fourier transform (DFT) of the audio frames $\mathbf{F}^a$ and then crops the center with a bounding box with the shape of (S, kN^a) where S refers to the dimension of the spectral feature to get $\tilde{\mathbf{F}}^a_{crop}$. Then the output of the inverse discrete Fourier transform (IDFT) $\hat{\mathbf{F}}^a$ is taken as the compressed audio, as shown below,

$$\begin{aligned} \tilde{\mathbf{F}}^a &= DFT(\mathbf{F}^a) \\ \tilde{\mathbf{F}}^a_{crop} &= \mathbf{F}^a(S, kN^a) \\ \hat{\mathbf{F}}^a &= IDFT(\tilde{\mathbf{F}}^a_{crop}). \end{aligned} \tag{2}$$

Down-sampling is performed on $\mathbf{F}^v$ to obtain compressed visual frames $\hat{\mathbf{F}}^v$,

$$\hat{\mathbf{F}}^v = \mathbf{F}^v(m/k), \qquad m = 0, 1, 2, ..., N_v - 1 \tag{3}$$

where k denotes the compression rate, ranging from 0 to 1.

We extract audio and visual features from compressed frames using PANNs and Inception-v3. Then, latent vectors are acquired with a mean operator. We employ knowledge distillation from the teacher network to increase the accuracy of caption generation. A neural network with two hidden layers is utilized to increase the resemblance of latent vectors to the teacher. The network is trained to minimize the L1 loss

TABLE I: Performance metric evaluation results on the MSR-VTT test set

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	SPICE	SCORE	Diff (%)
$\mathcal{L}_{rep}$ Student (k = 0.2)	0.722	0.555	0.422	0.311	0.267	0.236	0.554	0.045	0.321	0.127
$\mathcal{L}_{rep}$ Student (k = 0.4)	0.715	0.546	0.411	0.294	0.223	0.234	0.539	0.043	0.306	0.168
$\mathcal{L}_{rep}$ Student (k = 0.6)	0.709	0.542	0.412	0.300	0.232	0.231	0.543	0.041	0.308	0.163
$\mathcal{L}_{rep}$ Student (k = 0.8)	0.719	0.550	0.413	0.300	0.256	0.235	0.545	0.046	0.315	0.144
$\mathcal{L}_{rep} + \mathcal{L}_{CE}$ Student (k = 0.2)	0.766	0.613	0.476	0.357	0.375	0.256	0.585	0.054	0.365	0.008
$\mathcal{L}_{rep} + \mathcal{L}_{CE}$ Student (k = 0.4)	**0.774**	**0.618**	0.473	0.348	0.359	0.256	0.582	**0.055**	0.361	0.019
$\mathcal{L}_{rep} + \mathcal{L}_{CE}$ Student (k = 0.6)	0.769	0.616	0.478	0.357	0.375	**0.258**	**0.586**	**0.055**	0.366	0.005
$\mathcal{L}_{rep} + \mathcal{L}_{CE}$ Student (k = 0.8)	0.765	0.614	**0.479**	**0.358**	0.366	0.255	0.583	0.054	0.362	0.016
Teacher	0.760	0.612	0.473	0.352	**0.397**	0.254	0.583	0.054	**0.368**	0.000

between student and teacher latent vectors. We denote this loss as $\mathcal{L}_{rep}$ where rep refers to representation. We train the teacher network, and then the teacher guides the optimization of the parameters of the student network. In this study, we train the student-teacher network with the following losses:

$\mathcal{L}_{rep}$: The student network is only trained by the $\mathcal{L}_{rep}$ loss and is learned to mimic the audio-visual features of the teacher network. Then, the language model is trained with the updated neural network.

$\mathcal{L}_{rep} + \mathcal{L}_{CE}$: We employ both $\mathcal{L}_{rep}$ and $\mathcal{L}_{CE}$ losses to minimize the representation loss and maximize the captioning accuracy.

III. Experimental Evaluations

A. Setup and Performance Metrics

The proposed approach is evaluated on the MSR-VTT dataset [29], which initially consists of 10,000 videos, each with 20 ground-truth captions. However, by the time the experiments are executed, only 5,074 and 2,123 videos are available from the training and testing sets, respectively. Several performance metrics are employed to measure the accuracy of the video captioning approach, including metrics for evaluation of translation with explicit ordering (METEOR) [30], bilingual evaluation understudy (BLEU) [31], consensus-based image description evaluation (CIDEr) [32], and recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L) [33], and semantic propositional image caption evaluation (SPICE) [34].

The ranking of the results is based on a final SCORE which is calculated as an average of all performance metrics. In calculating the final SCORE, we used the mean of the BLEU scores. For the experiments, the visual frames of the videos are resized into the shape of 3×299×299. We utilized tokenization and punctuation removal on the ground-truth captions of the training set. The latent vector size of the layers in the language models is set to 2,576, and the dimension of the linear layer output is equal to the vocabulary length. We evaluated the proposed approach with $0.2, 0.4, 0.6,$ and 0.8 compression ratios.

TABLE II: Time consumption evaluation results on random 100 videos from the MSR-VTT test set

Network	average time consumption (s)	Diff (%)
Student (k = 0.2)	2.77	79.1
Student (k = 0.4)	5.65	57.4
Student (k = 0.6)	8.31	37.4
Student (k = 0.8)	11.03	16.9
Teacher	13.28	0.0

B. Results & Discussion

The accuracy and time consumption of the teacher and student networks are measured with the test set of the MSR-VTT dataset under the $\mathcal{L}_{rep}$, and $\mathcal{L}_{rep} + \mathcal{L}_{CE}$ losses. In the evaluations, we compressed the frames on the student networks to enable faster inference time. The results for the students and teacher networks are given in Table I, while time consumptions are shown in Table II.

Using only the $\mathcal{L}_{rep}$ loss resulted in poor captioning performance in all performance metrics regarding the teacher network, as seen in Table I. Notably, among the student networks trained with the $\mathcal{L}_{rep}$ loss, the compression rate of 0.2 has achieved the highest final SCORE. However, the combination of the $\mathcal{L}_{rep}$ and $\mathcal{L}_{CE}$ losses in the student networks offered an accuracy approaching the level of the teacher model across all performance metrics.

The captioning accuracy of the student network is increased from 0.321 to 0.365 with $\mathcal{L}_{rep}$+$\mathcal{L}_{CE}$ under the compression rate of 0.2. The difference between the accuracy of the teacher and student network dropped from 0.127% to 0.008%. However, the student network with a 0.4 compression rate leveraged the final SCORE from 0.306 to 0.361, which is still lower than that of the compression rate at 0.2. We achieved the highest final SCORE at 0.366 using the student network with a compression rate of 0.6. This is followed by the compressed student network with a compression rate of 0.8, with a final SCORE of 0.362. Furthermore, the student networks with compressed audio and visual frames scored higher across some metrics than the teacher. This indicates that student networks can generate accurate captions similar to the teacher. In Table

II, we present the time consumption of feature extraction for both audio and visual frames from randomly selected 100 videos from the test set of the MSR-VTT dataset. The compression rate 0.8 reduces feature extraction time up to 16.9%, while 0.6 compression rate decreases the audio-visual feature extraction time by about 37.4%. Similarly, 0.4 and 0.2 have reduced the inference time by 57.4% and 79.1%, respectively. Table II shows that the student networks reduce inference time significantly compared to the teacher network.

IV. Conclusion

In this study, we have presented a simple pooling front-end and down-sampling method to reduce the number of audio and visual frames in a video for video captioning. Furthermore, we have proposed a teacher-student based-network to leverage the accuracy of caption generation with knowledge distillation. We used $\mathcal{L}_{rep}$ representation and $\mathcal{L}_{CE}$ cross-entropy loss for network training. The proposed approach is evaluated on the MSR-VTT dataset. Experimental results show that the proposed approach significantly reduces the inference time with a negligible drop in captioning accuracy.

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 projects: 120N995, & 623805725) and by the scientific research projects coordination unit of Izmir Katip Celebi University (project no: 2021-ÖDL-MÜMF-0006, & 2022-TYL-FEBE-0012). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] Betül Uslu, Özkan Çaylı, Volkan Kılıç, and Aytuğ Onan, "Resnet based deep gated recurrent unit for image captioning on smartphone," *European Journal of Science and Technology*, , no. 35, pp. 610–615, 2022.

[2] Bengü Fetiler, Özkan Çaylı, Özge Taylan Moral, Volkan Kılıç, and Aytuğ Onan, "Video captioning based on multi-layer gated recurrent unit for smartphones," *European Journal of Science and Technology*, , no. 32, pp. 221–226, 2021.

[3] Rumeysa Keskin, Özkan Çaylı, Özge Taylan Moral, Volkan Kılıç, and Aytuğ Onan, "A benchmark for feature-injection architectures in image captioning," *European Journal of Science and Technology*, , no. 31, pp. 461–468, 2021.

[4] Xinhao Mei, Xubo Liu, Mark D Plumbley, and Wenwu Wang, "Automated audio captioning: An overview of recent progress and new challenges," *EURASIP journal on audio, speech, and music processing*, vol. 2022, no. 1, pp. 1–18, 2022.

[5] Jianyuan Sun, Xubo Liu, Xinhao Mei, Mark D Plumbley, Volkan Kilic, and Wenwu Wang, "Automated audio captioning via fusion of low-and high-dimensional features," *arXiv preprint arXiv:2210.05037*, 2022.

[6] Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H Tang, et al., "Visually-aware audio captioning with adaptive audio-visual attention," *arXiv preprint arXiv:2210.16428*, 2022.

[7] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[8] Arshdeep Singh and Mark D Plumbley, "A passive similarity based cnn filter pruning for efficient acoustic scene classification," *arXiv preprint arXiv:2203.15751*, 2022.

[9] Arshdeep Singh and Mark D Plumbley, "Low-complexity cnns for acoustic scene classification," *arXiv preprint arXiv:2207.11529*, 2022.

[10] Hayato Futami, Hirofumi Inaguma, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Distilling the knowledge of bert for sequence-to-sequence asr," *arXiv preprint arXiv:2008.03822*, 2020.

[11] Liang Lu, Michelle Guo, and Steve Renals, "Knowledge distillation for small-footprint highway networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4820–4824.

[12] Kwanghee Choi, Martin Kersner, Jacob Morton, and Buru Chang, "Temporal knowledge distillation for on-device audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 486–490.

[13] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[14] Jinhua Liang, Tao Zhang, and Guoqing Feng, "Channel compression: Rethinking information redundancy among channels in cnn architecture," *IEEE Access*, vol. 8, pp. 147265–147274, 2020.

[15] Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Mark D Plumbley, and Wenwu Wang, "Simple pooling front-ends for efficient audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[17] Xubo Liu, Xinhao Mei, Qiushi Huang, Jianyuan Sun, Jinzheng Zhao, Haohe Liu, Mark D Plumbley, Volkan Kilic, and Wenwu Wang, "Leveraging pre-trained bert for audio captioning," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1145–1149.

[18] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang, "Diverse audio captioning via adversarial training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8882–8886.

[19] Xinhao Mei, Qiushi Huang, Xubo Liu, Gengyun Chen, Jingqian Wu, Yusong Wu, Jinzheng Zhao, Shengchen Li, Tom Ko, H Lilian Tang, et al., "An encoder-decoder based audio captioning system with transfer and reinforcement learning," *arXiv preprint arXiv:2108.02752*, 2021.

[20] Pavel Ostyakov, Elizaveta Logacheva, Roman Suvorov, Vladimir Aliev, Gleb Sterkin, Oleg Khomenko, and Sergey I. Nikolenko, "Label denoising with large ensembles of heterogeneous neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 250–261.

[21] Rongcheng Lin, Jing Xiao, and Jianping Fan, "Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 1092–1101.

[22] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M Khapra, "Efficient video classification using fewer frames," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVF)*. 2019, pp. 354–363, IEEE.

[23] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[24] Özkan Çaylı, Burak Makav, Volkan Kılıç, and Aytuğ Onan, "Mobile application based automatic caption generation for visually impaired," in *International Conference on Intelligent and Fuzzy Systems (INFUS)*. IEEE, 2020, pp. 1532–1539.

[25] Özkan Çaylı, Volkan Kılıç, Aytuğ Onan, and Wenwu Wang, "Auxiliary classifier based residual rnn for image captioning," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1126–1130.

[26] Stefan Wager, Sida Wang, and Percy S Liang, "Dropout training as adaptive regularization," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[27] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVF)*. 2015, pp. 3128–3137, IEEE.

[28] Selman Aydın, Özkan Çaylı, Volkan Kılıç, and Aytuğ Onan, "Sequence-to-sequence video captioning with residual connected gated recurrent

units," *European Journal of Science and Technology*, , no. 35, pp. 380–386, 2022.

[29] Jun Xu, Tao Mei, Ting Yao, and Yong Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVF)*. 2016, pp. 5288–5296, IEEE.

[30] Alon Lavie and Abhaya Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVF)*. 2015, pp. 4566–4575, IEEE.

[33] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Association for Computational Linguistics (ACL) Workshop*, 2004, pp. 1–8.

[34] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 382–398.