Auxiliary Classifier based Residual RNN for Image Captioning

Özkan Çaylı[1], Volkan Kılıç[1*], Aytuğ Onan[2], Wenwu Wang[3‡]
[1]Electrical and Electronics Engineering Graduate Program, İzmir Katip Çelebi University, Turkey
[2]Department of Computer Engineering, İzmir Katip Çelebi University, Turkey
[3]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
Email: *volkan.kilic@ikcu.edu.tr; ‡w.wang@surrey.ac.uk

***Abstract*—Image captioning aims to generate a description of visual contents with natural language automatically. This is useful in several potential applications, such as image understanding and virtual assistants. With recent advances in deep neural networks, natural and semantic text generation has been improved in image captioning. However, maintaining the gradient flow between neurons in consecutive layers becomes challenging as the network gets deeper. In this paper, we propose to integrate an auxiliary classifier in the residual recurrent neural network, which enables the gradient flow to reach the bottom layers for enhanced caption generation. Experiments on the MSCOCO and VizWiz datasets demonstrate the advantage of our proposed approach over the state-of-the-art approaches in several performance metrics.**

***Index Terms*—image captioning, recurrent neural networks, residual connections, auxiliary classifier**

I. Introduction

Image captioning is a task of generating a meaningful and grammatically accurate sentence that describes the contents of an image. Recently, it has received increasing interest from computer vision, signal processing and natural language processing fields due to its potential applications in social media services, image indexing, and virtual assistance for the visually impaired [1]–[3].

Image captioning approaches mostly employ retrieval-based, template-based, and neural encoder-decoder frameworks [4]. In retrieval-based methods, the images that are similar to the input image are found in the training set, and a caption is selected from the reference captions. In template-based methods, a caption is generated from fixed templates by matching the visual information of the detected objects and actions. One issue of these approaches is that they heavily rely on the training set and reference captions. With recent advances in deep learning, the neural encoder-decoder framework has gained attention due to its advantage in extracting vision-language features [5]. This framework adopts two sub-networks, namely, an encoder and a decoder, where the encoder extracts features of an image as a latent vector, and then the decoder generates a caption word-by-word [6]. Conventionally, there are four architectures on where to put features in the decoder, which are init-inject, pre-inject, par-inject, and merge [7]. The init-inject architecture utilizes the features as the initial hidden state of a recurrent neural network (RNN). Whereas in pre-inject, the features are used as the first input of the decoder. In par-inject, the visual and language features are fed as the input to an RNN. In the merge architecture, however, the visual and language features are combined at the output of an RNN.

Training deep learning from scratch is a time-consuming process due to the presence of a large number of parameters in the neural network, which makes implementing deep learning in an encoder computationally expensive. Therefore, transfer learning is employed where pre-trained deep learning models are used for feature extraction, which reduces the computational load for optimizing the model parameters [8]–[10]. Vinyals et al. [9] propose a neural image caption generator with a GoogleNet [11] architecture which is trained on the ImageNet dataset [12]. Similarly, a region-based convolutional neural network (CNN) pre-trained on the ImageNet dataset is used to extract 19 objects and their positions as features [13]. In language modeling, a caption is generated as a sequence of words, for which conventional methods utilize an RNN for sequential data processing [14], [15].

However, propagating features through words can be limited by the gradient vanishing and exploding issues in simple RNNs. Therefore, gating mechanisms such as long-short term memory (LSTM) and gated recurrent units (GRU) are employed in RNNs [16], [17]. Donahue et al. [18] introduces an LSTM based decoder to exploit its advantage in processing long term sequences. In [16], the number of GRU layers is increased in language modeling for capturing complex data attributes. Dropout [19] and residual connections [20], [21] are employed to carry effective attributes through consecutive layers, resulting in a better convergence rate and improved accuracy. Furthermore, an auxiliary classifier is utilized in object detection and classification, which offers additional feedback to back-propagate the gradient information to the lower layers [21], [22].

Inspired by the use of auxiliary classifiers in object detection and localization [23], we are interested in studying whether it is also helpful in language modeling. To this end, we propose an auxiliary classifier based residual RNN in the decoder to improve image captioning. We performed the experiments on the MSCOCO [24] and VizWiz [25] datasets and evaluated the performance of the proposed method using performance metrics such as BLEU-n [26], METEOR [27], and CIDEr [28]. In addition, we apply a grid search on RNN types, GRU

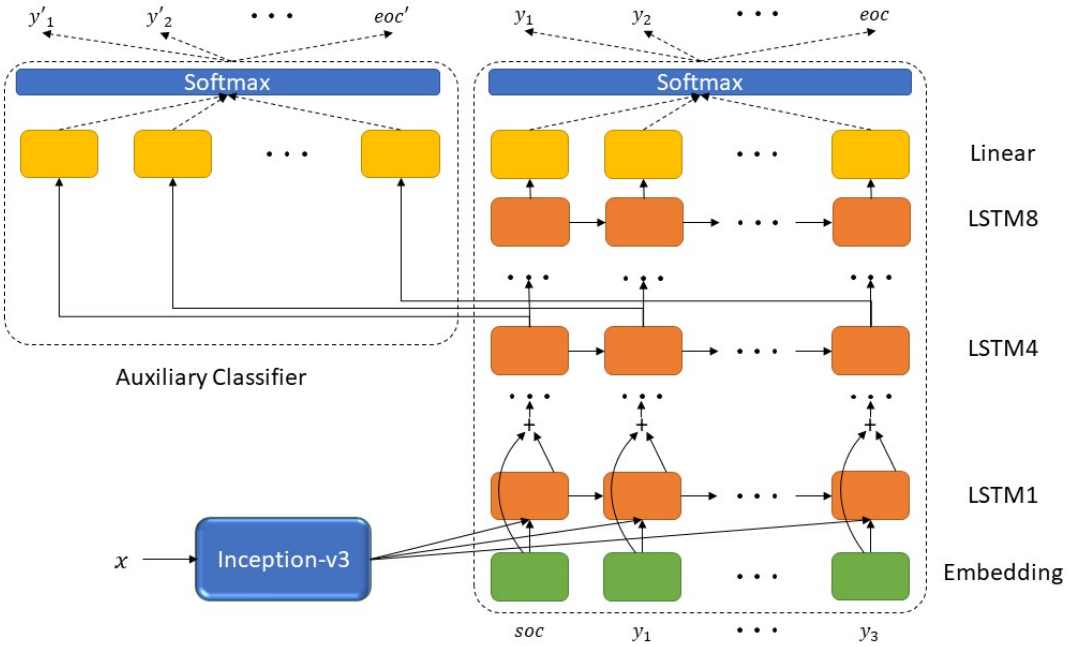

Fig. 1: The proposed approach

and LSTM, under only par-inject and pre-inject architectures. The merge architecture combines vision features and language features at the output, as a result, it is not suitable for testing with the auxiliary classifier. In addition, it was shown in [29] that pre-inject and par-inject offer better performance as compared with the init-inject architecture. We show in our experiments that the utilization of the auxiliary classifier has improved the captioning performance in both quantitative and qualitative evaluations.

The organization of the paper is presented as follows. Section II introduces the proposed image captioning approach in detail. Section III presents experimental evaluation, which consists of datasets, performance metrics, data preparation, results and discussions. Conclusions are drawn in Section IV.

II. Proposed Image Captioning Approach

In this section, the proposed image captioning approach is presented. We utilize Inception-v3 deep CNN architecture for feature extraction. The architecture takes an input image denoted as x and extracts features at the output of the global average pooling layer. The features are fed into a language model consisting of embedding, RNN, and linear layers. The purpose of the language model is to sequentially predict a word until an end-of-caption word (eoc) is generated. In the study, multi-layer GRU and LSTM are utilized in the model, which produces a hidden state for predicting the next word. GRU and LSTM compute the hidden state h_t as shown in Equations (1) and (2), respectively.

$$\begin{aligned} r_t &= \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \\ z_t &= \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \\ n_t &= tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{(t-1)} + b_{hn})) \\ h_t &= (1 - z_t) \odot n_t + z_t \odot h_{(t-1)} \end{aligned} \tag{1}$$

where r_t, z_t, and n_t are the reset, update, and new gates of GRU, respectively. We refer to the Hadamard product as $\odot$ and the sigmoid activation function as σ.

$$\begin{aligned} p_t &= \sigma(W_{ip}x_t + b_{ip} + W_{hp}h_{(t-1)} + b_{hp}) \\ f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}) \\ g_t &= tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}) \\ o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \\ c_t &= f_t \odot c_{t-1} + p_t \odot g_t \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{2}$$

where p_t, f_t, g_t, and o_t are the input, forget, cell, and output gates of LSTM, respectively. In the above equations, weights (W) and biases (b) each have two subscripts, where the first one either refers to input i or hidden state h, while the second one indicates the corresponding gate. For example, W_{ip} is the weight in gate p for the input and b_{hg} is the bias in gate g for the hidden state h.

We follow the pre-inject and par-inject architectures described in [7] to feed the features to the language model. In pre-inject, image features extracted in the encoder are taken

TABLE I: Performance metric results on validation sets

		GRU		LSTM	
		Pre	**Par**	**Pre**	**Par**
VizWiz	**Aux**	0.235	0.185	0.262	**0.262**
	NoAux	0.214	0.173	0.260	0.260
MSCOCO	**Aux**	0.746	0.743	0.809	**0.820**
	NoAux	0.745	0.720	0.796	0.800

as the first input of RNN. The vectors that are generated in the embedding layer by converting the words are taken as the following inputs for the RNN. However, in par-inject, image features are concatenated with word vectors as the inputs of the RNN. Furthermore, residual connections are utilized to carry the information through consecutive layers additively as shown in Equation (3). The m_t^l is the output of the l-th layer, h_t^l is the hidden state, and x_t^l is the input to the $(l+1)$-th layer of RNN at the t-th time.

$$\begin{aligned} m_t^l, h_t^l &= RNN_l(x_t^{l-1}, h_{t-1}^l) \\ x_t^l &= m_t^l + x_t^{l-1} \\ m_t^{l+1}, h_t^{l+1} &= RNN_{l+1}(x_t^l, h_{t-1}^{l+1}) \end{aligned} \tag{3}$$

A dropout is applied between layers to keep only useful information. The output of the fourth RNN layer is additionally fed to an auxiliary classifier which consists of a linear layer, as shown in Fig. 1. We additionally calculate loss and gradients for the auxiliary classifier to back-propagate them to the lower layers. A target sentence is referred to as $\hat{Y} = \hat{y}_1, \hat{y}_2, ..., \hat{y}_N$ where $\hat{y}_n$ corresponds to the n-th word. Similarly Y and $Y^{'}$ are the sequential predictions of the network and the auxiliary classifier, respectively. The cross-entropy (CE) loss is utilized as the criterion for training and the total loss is calculated as: $loss = a * CE(Y^{'}, \hat{Y}) + CE(Y, \hat{Y})$ where a is a parameter for weighting the auxiliary classifier.

III. Experimental Evaluations

This section presents experimental evaluations of the proposed approaches on MSCOCO [24] and VizWiz [25] datasets. First, the datasets, performance metrics and experimental setup are introduced. Then, comparative results between the proposed and existing approaches are given and discussed.

A. Setup and Performance Metrics

The proposed approach is evaluated on the MSCOCO [24] and VizWiz [25] datasets. The MSCOCO dataset contains 123,287 training and validation images, with five ground-truth (GT) captions for each image. Similarly, the VizWiz dataset consists of 31,704 images with each annotated with five captions. Specifically, the images in the VizWiz dataset were captured by visually impaired people, thus, mainly concerning indoor objects and activities.

A number of metrics are proposed in the literature to measure the performance of the image captioning approaches, including bilingual evaluation understudy (BLEU) [26], metric for evaluation of translation with explicit ordering (METEOR) [27], recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L) [30], semantic propositional image caption evaluation (SPICE) [31], and consensus-based image description evaluation (CIDEr) [28]. Among these metrics, CIDEr is a better metric for evaluating image captioning performance as compared with others which are originally derived for machine translation, resulting in a default comparison metric in the MSCOCO evaluation server. Therefore, the results in this paper have been sorted based on the CIDEr metric.

TABLE II: Online evaluation results on VizWiz test set

BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	METEOR	CIDEr	SPICE
0.577	0.379	0.240	0.141	0.401	0.155	0.303	0.103

In our evaluation, the images were first resized to 3-by-299-by-299 before being used in the Inception-v3 for feature extraction. Preprocessing has been employed on the training set to remove punctuations and tokenize the reference captions. We selected captions between 6-15 words to maintain consistency with paddings. We removed the captions describing the "quality issues" phrase such as "Quality issues are too severe to recognize a content" and "Focused well, but the image has some drawbacks and quality issues" from the training set of the VizWiz as it may result in misleading captions. We set latent vector size in embedding, RNNs, and linear layer to 2,048. The output of the network is adjusted to the vocabulary size of the dataset. A stochastic gradient descent algorithm was used to optimize the network with a learning rate of 0.01, and the weight for the auxiliary classifier is set to 0.3.

B. Results & Discussion

For qualitative and quantitative analysis of the proposed approach, evaluations were performed in several steps, including offline (in the workstation) and online evaluations in MSCOCO and VizWiz servers.

1) Offline evaluation: First, the performance of the proposed approach was tested under two types of injection (pre and par) architectures with GRU and LSTM networks in the case of auxiliary and without auxiliary classifiers. Table I shows the results in terms of the CIDEr metric obtained on the validation sets of both MSCOCO and VizWiz datasets. In the experiments, we employed an auxiliary classifier that provides additional feedback to back-propagate the gradient information to the lower layers, leading to an efficient weight update. We found that the proposed model achieves higher accuracy with an auxiliary classifier due to the back-propagation of the gradient information preventing vanishing gradient issues in lower layers. In addition, LSTM performs better under par-inject architecture, while GRU achieves a higher score with pre-inject on both MSCOCO and VizWiz datasets regardless of using the auxiliary classifier. Further, LSTM yielded better results than GRU for all cases.

2) Online Evaluation: According to the results given in Table I, the auxiliary classifier based residual LSTM with

TABLE III: Online evaluation results on MSCOCO test set

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	METEOR	CIDEr
Chen et al. [32]	0.268	0.142	0.082	0.049	0.248	0.110	0.398
Karpathy et al. [13]	0.625	0.45	0.321	0.230	-	0.195	0.660
You et al. [33]	0.510	0.330	0.219	0.148	0.394	0.170	0.701
Proposed	**0.668**	**0.490**	**0.346**	**0.242**	**0.491**	0.222	0.747
Human [24]	0.663	0.469	0.321	0.217	0.484	**0.252**	**0.854**

TABLE IV: Sample images from MSCOCO (first two columns) and VizWiz (last two columns) with ground-truth and generated captions.

			
GT: A woman riding skis down a snow covered slope.	**GT:** Single train parked at a train station on a clear day.	**GT:** A woman is holding a plastic bottle with a green liquid in it.	**GT:** Big clear empty plastic bottle face down with food label showing
GT: A woman skiing down a snowy, tree-lined path.	**GT:** A train is near an empty outdoor station.	**GT:** A green bottle of liquid with black writing on it.	**GT:** Back of a large, empty Ocean Spray Cranberry Juice jug that shows the writing on the label.
GT: a woman is wearing a pink and black jacket is sking	**GT:** there is a yellow and red train at a train stop	**GT:** Green bottle of something, but image is too blurry to read.	**GT:** an empty bottle of ocean spray juice on a speckled tan surface.
GT: A woman on skis sliding down a snowy hill.	**GT:** A yellow an red colored train that is approaching a pickup/drop off point.	**GT:** A bottle of some green liquid is in a palm above a white table.	**GT:** An empty bottle of Ocean Spray juice rests faced-down on a white surface.
GT: A woman skiing on a path by some trees.	**GT:** A yellow and red train riding into a station.	**GT:** The front label of a plastic bottle of food.	**GT:** The back of a nearly empty bottle of Ocean Spray cranberry juice.
With Aux: a person riding skis down a snow covered slope	**With Aux:** a train traveling down a track next to a forest	**With Aux:** a bottle of green juice is held up by a person's hand	**With Aux:** a bottle of liquid is on a counter top
Without Aux: a man riding skis down a snow covered slope	**Without Aux:** a train is going down the tracks in the middle of the road	**Without Aux:** a bottle of water is on a table	**Without Aux:** a bottle of kraft juice is on a table

par-inject has the highest CIDEr score across both datasets. Here, we have further tested the proposed approach in the online server of VizWiz and MSCOCO datasets. Table II demonstrates the test results on the official VizWiz evaluation server. The approach took fifth place on the leaderboard in CIDEr metric order. Similarly, Table III presents the results of the proposed approach on the MSCOCO evaluation server. Additionally, we made comparison with state-of-the-art approaches, including [13], [32], [33], and human agreement scores in [24].

The proposed approach outperforms others in terms of BLEU-1, BLEU-2, BLEU-3, BLEU-4, and Rouge-L metrics. However, it is worth noting that the performance of the proposed approach is slightly lower than Human [24], only in METEOR and CIDEr metrics.

3) Qualitative Analysis: In Table IV, we present four images with their GT and generated captions. The images have been selected from the MSCOCO and VizWiz validation sets in the first and last two columns, respectively. In the first column of Table IV, "man" and "person" is the only difference between generated captions. For the GT captions, the person is a woman. Therefore, the word "person" is more accurate than "man". In the second column, the network with auxiliary classifier (denoted as "with aux") correctly captures the train and forest in the background into a grammatically accurate caption. In the third column, "with aux" describes all the aspects of the image in a syntactically correct caption. On the other hand, in the network without an auxiliary classifier (denoted as "without aux"), just the word "bottle" was predicted correctly. Finally, the generated captions "with aux" containing the phrases "liquid" and "counter tap" are

more accurate than their counterpart "kraft juice" and "table" generated by "without aux" in the fourth column.

The study reveals that the auxiliary classifier based residual RNN with LSTM and par-inject achieves higher performance metric scores and generates more accurate captions.

IV. Conclusion

In this paper, we have presented an auxiliary classifier based RNN decoder for image captioning using residual connections in multi-layer LSTM under par-inject architecture. The auxiliary classifier back-propagates the gradient information to the lower layers while residual connections carry effective attributes through multi-layer LSTM, leading to modulation of detailed contextual information from the image. The proposed approach has been evaluated on VizWiz and MSCOCO datasets and compared with the state-of-the-art approaches. Offline and online evaluation results demonstrated that the proposed approach generates grammatically and semantically more accurate captions compared to its counterparts.

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). 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] B. Makav and V. Kılıç, "Smartphone-based image captioning for visually and hearing impaired," in *11th International Conference on Electrical and Electronics Engineering*. IEEE, 2019, pp. 950–953.

[2] S. Aydın, Ö. Çaylı, V. Kılıç, and A. Onan, "Sequence-to-sequence video captioning with residual connected gated recurrent units," *Avrupa Bilim ve Teknoloji Dergisi*, no. 35, pp. 380–386, 2022.

[3] B. Uslu, Ö. Çaylı, V. Kılıç, and A. Onan, "Resnet based deep gated recurrent unit for image captioning on smartphone," *Avrupa Bilim ve Teknoloji Dergisi*, no. 35, pp. 610–615, 2022.

[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European Conference on Computer Vision*. Springer, 2010, pp. 15–29.

[5] B. Makav and V. Kılıç, "A new image captioning approach for visually impaired people," in *11th International Conference on Electrical and Electronics Engineering*. IEEE, 2019, pp. 945–949.

[6] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[7] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Natural Language Engineering*, vol. 24, no. 3, pp. 467–489, 2018.

[8] Ö. Çaylı, B. Makav, V. Kılıç, and A. Onan, "Mobile application based automatic caption generation for visually impaired," in *International Conference on Intelligent and Fuzzy Systems*, 2020, pp. 1532–1539.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[10] R. Keskin, Ö. Çaylı, Ö. T. Moral, V. Kılıç, and A. Onan, "A benchmark for feature-injection architectures in image captioning," *European Journal of Science and Technology*, no. 31, pp. 461–468, 2021.

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

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[13] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[14] R. Keskin, Ö. T. Moral, V. Kılıç, and A. Onan, "Multi-gru based automated image captioning for smartphones," in *29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2021, pp. 1–4.

[15] M. Baran, Ö. T. Moral, and V. Kılıç, "Akıllı telefonlar için birleştirme modeli tabanlı görüntü altyazılama," *Avrupa Bilim ve Teknoloji Dergisi*, no. 26, pp. 191–196, 2021.

[16] V. Kılıç, "Deep gated recurrent unit for smartphone-based image captioning," *Sakarya University Journal of Computer and Information Sciences*, vol. 4, no. 2, pp. 181–191, 2021.

[17] B. Fetiler, Ö. Çaylı, Ö. T. Moral, V. Kılıç, and A. Onan, "Video captioning based on multi-layer gated recurrent unit for smartphones," *European Journal of Science and Technology*, no. 32, pp. 221–226, 2021.

[18] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

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

[20] J.-H. Luo and J. Wu, "Neural network pruning with residual-connections and limited-data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1458–1467.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *The 31st AAAI Conference on Artificial Intelligence*, 2017.

[22] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[24] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[25] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *European Conference on Computer Vision*. Springer, 2020, pp. 417–434.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[27] A. Lavie and A. 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.

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

[29] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894–4902.

[30] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop*. ACL, 2004, pp. 1–8.

[31] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[32] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin, and J. Luo, ""factual"or"emotional": Stylized image captioning with adaptive learning and attention," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 519–535.

[33] Q. You, H. Jin, and J. Luo, "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions," *arXiv preprint arXiv:1801.10121*, 2018.