A transformer-based deep learning model for evaluation of accessibility of image descriptions

Raju Shrestha
raju.shrestha@oslomet.no
Department of Computer Science, Oslo Metropolitan University (OsloMet)
Oslo, Norway

ABSTRACT

Images have become an integral part of digital and online media and they are used for creative expression and dissemination of knowledge. To address image accessibility challenges to the visually impaired community, adequate textual image descriptions or captions are provided, which can be read through screen readers. These descriptions could be either human-authored or software-generated. It is found that most of the image descriptions provided tend to be generic, inadequate, and often unreliable making them inaccessible. There are tools, methods, and metrics used to evaluate the quality of the generated text, but almost all of them are word-similarity-based and generic. There are standard guidelines such as NCAM image accessibility guidelines to help write accessible image descriptions. However, web content developers and authors do not seem to use them much, possibly due to the lack of knowledge, undermining the importance of accessibility coupled with complexity and difficulty understanding the guidelines. To our knowledge, none of the quality evaluation techniques take into account accessibility aspects. To address this, a deep learning model based on the transformer, a most recent and most effective architecture used in natural language processing, which measures compliance of the given image description to ten NCAM guidelines, is proposed. The experimental results confirm the effectiveness of the proposed model. This work could contribute to the growing research towards accessible images not only on the web but also on all digital devices.

CCS CONCEPTS

- **Computing methodologies → Neural networks**.

KEYWORDS

Image accessibility, Image description, Evaluation, NCAM guidelines, Neural networks, Deep learning, Transformer

ICMLC '22, February 18–21, 2022, Guangzhou, China
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9570-0. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Raju Shrestha. 2022. A transformer-based deep learning model for evaluation of accessibility of image descriptions. In *International Conference on Machine Learning and Computing, February 18–21, 2022, Guangzhou, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Images or photos are one of the most commonly used contents in this digital era. They not only remind us of people, places, feelings, and stories but also give an impression about them. Because of that people frequently take pictures, and publish them on the web or share them via social networks, thanks to the proliferation of Internet technology, smartphones, camera technology, and social media. Since an image can convey many ideas and information, it is said that 'a picture is worth a thousand words'. A person with normal vision can acquire this information by viewing the image. However, this is not possible for people with visual impairment, posing serious accessibility challenges [21]. In order to help make images accessible to the visually impaired community, Web Content Accessibility Guidelines (WCAG) version 2.1[1] recommends providing image descriptions or captions as alternative text (ALT text) so that they can be read for the user by assistive technologies such as screen reader. Accessible images help in ongoing efforts in bringing down the digital divide.

There are a massive number of images on the Internet which do not provide ALT text. And, the majority of them which has ALT text are either not accurate or not adequate to convey proper information [4, 15–17]. The rate of availability of ALT text in social media is even much lower [16]. On the other hand, simply providing image descriptions may not be valuable if those descriptions are not understandable or confusing. These have negative consequences for the visually impaired community as it creates obstacles for the users to staying socially connected [21].

In order to help write accessible image descriptions, there are standard guidelines such as WCAG 2.1 and NCAM (National Center for Accessible Media) guidelines[2]. WCAG 2.1 includes a broad range of recommendations for making web content more accessible to a wide range of people with disabilities including low and no vision (blind). NCAM provides guidelines for accessible media including different types of images such as maps, graphs, and general images. It also

[1] WCAG 2.1: https://www.w3.org/TR/WCAG21/
[2] NCAM Guidelines: http://diagramcenter.org/table-of-contents-2.html

provides a tool, called POET Training Tool [3] to learn and create accessible image descriptions.

Images could either be described to provide information about the image's visual features, or captioned to relate the image to the broader context. Detailed image descriptions can better address accessibility challenges [18]. In this paper, we use the terms caption and description interchangeably as we intend both to be accessible.

Most often image descriptions are written manually by web content developers or the owner of the images themselves. This could be a tedious and time-consuming process. Moreover, the lack of knowledge about accessibility guidelines could result in poor quality image descriptions in terms of accessibility. Dahal and Shrestha [11] proposed a method of writing accessible image descriptions based on NCAM guidelines by providing sample example cues. The method has shown to be useful for users who have no or minimal knowledge about image accessibility guidelines. Recent advances in machine learning and deep learning have led to an increasing number of models that generate text from images [2, 7, 20]. There are methods and tools proposed which can generate image descriptions automatically using software, such as VizWiz [6] and queried image description and free response image description [23]. However, image descriptions generated with these models tend to be generic, often unreliable, and inaccessible [7, 13].

Irrespective of the method used, generated image descriptions should be evaluated to make sure they are of good quality, i.e., accessible in the problem context here. This could be done by image accessibility experts. However, the manual method is tedious, time-consuming, error-prone, and may not even be practicable when there is a huge number of images. An alternative solution could be a software-based solution that can do the evaluation automatically. To our knowledge, only limited research has been done in this direction.

Bigham [5] proposed a classifier to automatically measure the quality of a given ALT text on a web page. The classifier uses various features such as similarity of the alternative text to the content of the page where it is used, alternative text that is known to be good or bad, and features of the image that it is describing. The major limitations of the study were that the test was carried out on a small dataset and it simply outputs the input ALT text as appropriate or inappropriate, without taking into account any standard accessibility guideline. Moreover, it is not intended to evaluate image descriptions of independent images which are not on a web page but are on digital devices or on social media.

Generated image descriptions are commonly evaluated using rule-based metrics such as BLEU [24], METEOR [3], ROUGE [19], or CIDEr [27]. Since these metrics mainly measure the word overlap between generated and reference text, they fail to correlate well with human judgments. SPICE [1] metric measures similarity of scene graphs constructed from the candidate and reference texts, showing better correlation with human judgments, but it fails to capture the syntactic structure of a sentence. To address those limitations, Cui et al. [10] proposed a Generative Adversarial Networks (GAN) based discriminative learning model, which was trained to distinguish between human and machine-generated image descriptions. However, none of these metrics take into account any accessibility guideline.

We proposed a neural network based machine learning model and framework for an automatic evaluation of accessibility of image descriptions using NCAM guidelines [25]. The model was based on a manual selection of features. Therefore, model performance obviously depends on the selected features. Thanks to deep learning, which mitigates manual selection of features as it can extract features automatically during training of the model. Because of this, deep learning has gained tremendous popularity and use recently. In this paper, a novel transformer [26] based deep learning model, which can automatically evaluate the quality of a given image description in terms of compliance to the 10 NCAM image accessibility guidelines, is proposed. The transformer is the most recent revolutionary deep learning architecture used in natural language processing (NLP). Unlike traditional recurrent neural network (RNN) and long-short-term memory (LSTM), a transformer has extremely long-term memory and enables parallel computation making it more efficient and effective, thanks to its special design of positional encoding, self-attention mechanism, and encoder-decoder architecture. The results show that the proposed model works very well in evaluating the accessibility of a given image description. We believe that this work contributes to the growing research towards accessible images.

The rest of the paper is organized a sfollows. Section 2 brifly describes the NCAM guidelines used. Section 3 presents the proposed model. Section 4 describes the experimental setup. The results are presented and discussed in Section 5. Finally, Section 6 concludes the paper.

2 NCAM IMAGE ACCESSIBILITY GUIDELINES

NCAM provides guidelines for the accessibility of almost all types of images including maps, graphs, and natural images. Among the fourteen guidelines listed in [11], ten guidelines which include eight guidelines common to all types of images and two guidelines specific to natural images are used in this study. Guidelines for the map and graph images are excluded because of the unavailability of the datasets with those types of images. A summarized list of these ten guidelines is given below.

1. The description should be succinct.
2. Colors should not be specified unless it is significant.
3. The new concept or terms should not be introduced.
4. The description should be started with a high-level context and drilled down to details to enhance understanding.
5. The active verbs in the present tense should be used.
6. Spelling, grammar, and punctuation should be correct.
7. Symbols should be written out properly.

[3] POET Training Tool: https://poet.diagramcenter.org

8. The description vocabulary should be added which adds meaning, for example, “map” instead of an image.
9. Physical appearance and actions should be explained rather than emotions and possible intentions.
10. The material should not be interpreted or analyzed; instead, the reader should be allowed to form their own opinions.

3 PROPOSED MODEL

The proposed transformer-based deep learning model, which classifies a given input image description as compliant or non-compliant to an NCAM guideline, consists of an encoder and a classifier as shown in Figure 1.

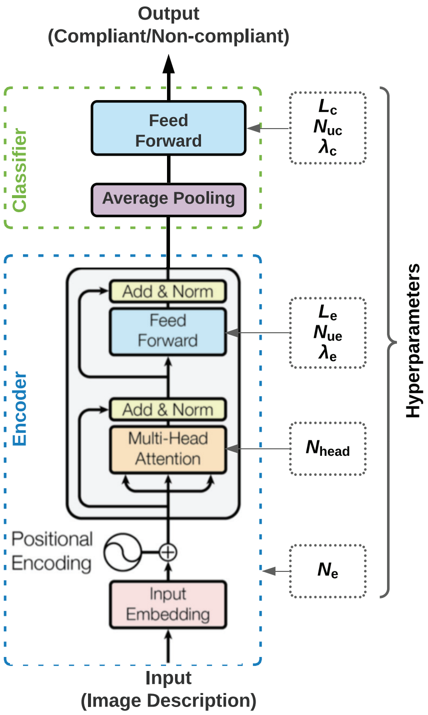

Figure 1: Architecture of the proposed model (Adapted from [26]), and the hyperparameters used in different components of the model.

The encoder is a transformer that maps an input image description into an abstract continuous representation that holds the learned information from the description. As in the original transformer encoder [26], the encoder is made up of input embedding, positional encoding, multi-head attention, and feed-forward networks. The output of each multi-head attention module and the feed-forward network is added to the residual connection and then passed through a layer normalization. Unlike in the original transformer encoder, which has one dense layer in the feed-forward network, the proposed model is made more generic to have more than one dense layer. In addition to the dropout layers in the multihead and dense layers in the feed-forward network, L1-regularization is introduced to overcome overfitting and for stable training of the models. Optimal values of the number of heads in the multi-head attention, number of encoder layers, number of units in these layers, and regularization parameters, are determined through a hyperparameter optimization (HPO) process.

The classifier uses the output from the encoder to classify the given input description as compliant or non-compliant to an NCAM image accessibility guideline. The output from the encoder is globally averaged and then passed through a feed-forward network, which finally outputs the prediction result from the model. Just like in the encoder, the output from each dense layer in the feed-forward network is passed through a dropout layer and subject to regularization. The optimal number of dense layers, the number of units in these layers, and regularization parameter values are determined through the HPO process.

In both the encoder and the classifier, outputs from the dense layers are batch normalized. Exponential Linear Unit (ELU) [9] is used as an activation function as it fixes some of the problems with Rectified Linear Unit (ReLU) such as dead ReLU. ELU helps the network nudge weights and biases in the right directions by producing small negative values instead of zero values.

Altogether there were fifteen hyperparameters defined in the model. Eight of them, number of encoder layers (N_e), number of heads in the multi-head attention in the encoder (N_head), number of layers (L_e), number of units in each layer (N_ue) and regularization parameter (λ_e) in the feedforward network of the encoder, number of layers (L_c), number of units in each layer (N_uc) and regularization parameter (λ_c) in the feed-forward network of the classifier were optimized through HPO process. Fixed values for encoder model dimension of 64, max positional encoding of 100, dropout rates of 0.1 (in multihead and dense layers in both the encoder and the classifier), a batch size of 32, and Adam optimizer with default learning rate are used as they produced good results in most cases.

A classifier model corresponding to each of the 10 NCAM guidelines is created by optimizing the hyperparameters through the HPO process. The optimal hyperparameters values used are given below in Table 1, Section 5.

4 EXPERIMENTAL SETUP

Ten models created for the 10 NCAM guidelines were trained, validated, tested, and evaluated. Experiments were conducted by implementing the models in Python 3 using Tensorflow 2 and running the code in Google Colab. HPO was performed in RayTune[4] using Baysian optimization technique with maximum validation F1 metric. For efficient training and also to address potential overfitting, early stopping was incorporated while training so that the training iteration stops when there is no more improvement after the last four iterations/epochs.

[4] Ray Tune: https://docs.ray.io/en/master/tune/index.html

The dataset and evaluation metrics used in the experiments are described next.

Dataset: The labeled dataset from [14], which were created from the popular Flickr8K dataset[5] containing 8K images, each image paired with 5 sets of image descriptions, is used. One set of image descriptions are manually labeled with their percentage compliance to the 10 NCAM guidelines by experts with a good knowledge of image accessibility and the guidelines. In this work, the percentage compliance values were converted to binary labels or classes, considering all above 50% compliance values as compliant (1) and equal or below 50% compliance values as non-compliant (0). Figure 2 shows six sample example images from the dataset along with their descriptions and compliance labels.

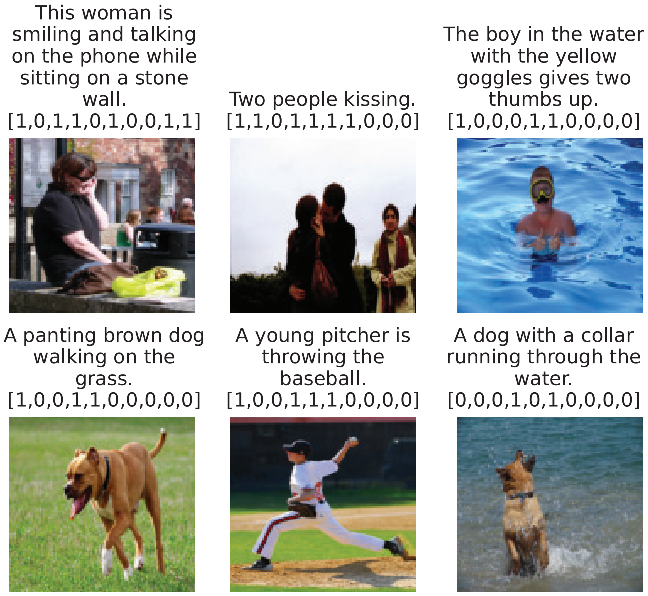
This woman is smiling and talking on the phone while sitting on a stone wall.
[1,0,1,1,0,1,0,0,1,1]
Two people kissing.
[1,1,0,1,1,1,1,0,0,0]
The boy in the water with the yellow goggles gives two thumbs up.
[1,0,0,0,1,1,0,0,0,0]
A panting brown dog walking on the grass.
[1,0,0,1,1,0,0,0,0,0]
A young pitcher is throwing the baseball.
[1,0,0,1,1,1,0,0,0,0]
A dog with a collar running through the water.
[0,0,0,1,0,1,0,0,0,0]

Figure 2: Sample example images from Flickr8K dataset with their image descriptions and compliance labels for the 10 NCAM guidelines.

The dataset was randomly split into training, validation, and test sets in the ratio of 70:15:15. As the dataset was not balanced, data splitting was done such that each class is distributed to the 3 sets in the same ratios.

Evaluation metrics: Since dataset is unbalanced, accuracy metric may not reflect a true performance of a model. Therefore, Precision and Recall metrics are used to evaluate the performance of the models as they are effective in the case of unbalanced datasets. Here, Precision can be defined as a fraction of correctly predicted compliant instances out of predicted compliant instances. Recall can be defined as a fraction of correctly predicted compliant instances out of actual compliant instances.

F1 score, which conveys the balance between the precision and the recall in a single metric value is also used. A scale-invariant and classification-threshold-invariant metric, area under the precision-recall curve (AUC-PR) is also used to evaluate model performance. An AUC measures the discriminating capability of a classifier to distinguish between compliant and non-compliant descriptions [22].

5 RESULTS AND DISCUSSION

The optimal hyperparameter values obtained from the HPO processes for the 10 models are given in Table 1. As anticipated, different optimal hyperparameter values are picked for the different models/classifiers. Training of the models showed good convergence of all the 10 models by 15 epochs.

Table 1: Optmimal hyperparameter values used in the 10 models.

Hyper-parameter	Model #									
	1	2	3	4	5	6	7	8	9	10
N_e	2	1	2	2	2	1	1	4	1	2
N_{head}	32	4	8	8	4	8	32	8	8	1
L_e	1	2	8	2	1	4	8	2	2	4
N_{ue}	32	8	32	16	8	32	8	8	16	16
λ_e	0.001	0.073	0.002	0.05	0.018	0.007	0.003	0.008	0.018	0.006
L_c	16	2	2	1	1	1	4	1	4	2
N_{uc}	16	32	16	8	8	16	16	512	32	16
λ_c	0.001	0.019	0.058	0.015	0.001	0.021	0.041	0.001	0.027	0.053

Performance metric values resulting from the tests of the models using the test data set are given in Table 2. Results were obtained with the prediction using the default threshold value of 0.5 (i.e., predicting probabilities higher than 0.5 as compliant and lower and equal to 0.5 as non-compliant). The results show that on average all the metric (precision, recall, and F1-score) values are around 0.97, which is very good. Looking at the individual models, the metric values are above the average values in almost all the models, indicating that the models were able to predict very well. Models #1 and #4 performed relatively poorer with precision, recall, and F1-score of around 0.9. The lower metric values in those models could be because the corresponding guidelines are relatively vague and difficult to interpret, therefore, compliance scores could become more subjective as human judgment could vary and become inconsistent. However, precision, recall, and F1-score of 0.9 could still be considered very well.

Results are also given in the form of precision-recall curves in Figure 3. We see that the average AUC is 0.99. This means that there is more than 99% chances that the models will be able to distinguish between compliant and non-compliant descriptions, which is outstanding [22].

[5] Flickr8KDataset:https://www.kaggle.com/adityajn105/flickr8k

Table 2: Test results of the models in terms of the four performance evaluation metrics.

Metric	Model #										Average
	1	2	3	4	5	6	7	8	9	10	
Precision	0.90	0.99	1.00	0.90	0.95	0.97	0.98	1.00	0.98	0.99	**0.97**
Recall	0.96	1.00	1.00	0.89	0.97	0.99	0.99	1.00	0.98	1.00	**0.98**
F1-Score	0.93	0.99	1.00	0.90	0.96	0.98	0.99	1.00	0.98	1.00	**0.97**
AUC-PR	0.96	1.00	1.00	0.96	0.98	0.99	0.99	1.00	0.99	1.00	**0.99**

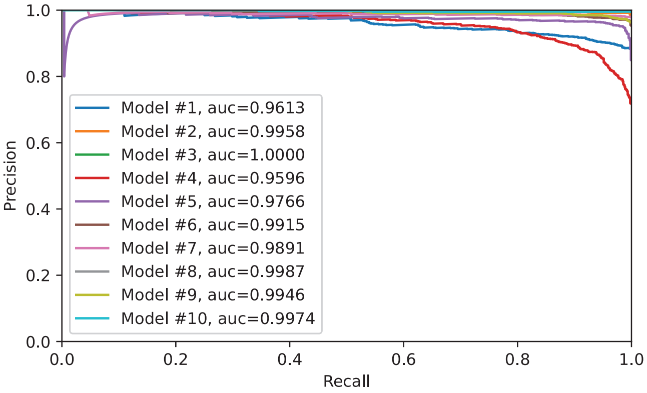

Figure 3: Precision-Recall curves. AUC values are shown along with the legends.

The experimental results thus show that the proposed model works very well in evaluating image descriptions in terms of compliance with the NCAM guidelines.

To our knowledge, this is the first and novel work on an automatic evaluation of accessibility of image descriptions based on compliance to NCAM guidelines. Before, this we proposed a neural network based model and framework which use manually selected features and predicts percentage ompliance level [25]. However, it used error and accuracy metrics instead of prediction, recall, F1-score and AUC metrics.

There are some shortcomings and challenges involved. First, the accessibility evaluation is purely based on the given image descriptions. It doesn't take into account the image and the context there but is not covered by the description. But it is assumed that this is the task of the author or generator of the image descriptions. Second, manually labeling percentage compliance of a large number of image descriptions is a tedious process and subject to individual judgment. Therefore, a relatively smaller Flickr8k dataset was used. Moreover, this dataset has mostly short, partial, and generic image descriptions. This study could be extended further with a larger dataset with detailed descriptions. One potential dataset could be a Wikipedia-based corpus proposed by Kreiss et al. [18], called Concadia, which consists of 96,918 images with descriptions, captions, and surrounding context and distinguishes between image descriptions and captions. This dataset could be used to further study in the future to analyze the quality of the captions and descriptions in terms of image accessibility.

It is worth mentioning here that a similar approach and process was used to test models based on RNN and LSTM as well but the results were significantly worse in most of the cases. This confirms the superiority of the transformer architecture over the traditional RNN and LSTM architectures [26, 28] in the context of the problem of interest in this work as well.

Based on the Transformer model, innovative, powerful, and scalable architectures such as BERT (Bidirectional Encoder Representations from Transformers [12]) and GPT (Generative Pre-trained Transformer, GPT-3 [8] is the most recent version) have been built, which are pre-trained on a wide range of data, in billions. These models are fine-tuned and used in NLP applications. In this work, we came up with transformer-based custom models trained from scratch for optimal size and performance. As a future work, it'd be interesting to see those pre-trained models for their performance and scalability.

6 CONCLUSION

The proposed transformer-based deep learning model, trained and tested with the Flickr8k dataset, has shown to perform excellently in evaluating accessibility of image descriptions in terms of their compliance to the 10 NCAM image accessibility guidelines. The model could be helpful to the web content developers and even general users to get an instant report of accessibility of the image descriptions they entered. This in turn will help towards more accessible images in the digital world.

In the future, the model could be extended with larger datasets containing more detailed image descriptions.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 382–398.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation withImproved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translationand/or Summarization.* Association for Computational Linguistics (ACL), 65–72.

[4] Ramayah Bavani, Azizah Jaafar, and Noor Faezah Mohd Yatim. 2010. A study on web experience among visually impaired users in Malaysia. In *2010 International Conference on User Science and Engineering (i-USEr)*. 11–15. https://doi.org/10.1109/IUSER.2010.5716714

[5] Jeffrey P. Bigham. 2007. Increasing Web Accessibility by Automatically Judging Alternative Text Quality. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (Honolulu, Hawaii, USA) *(IUI '07)*. Association for Computing Machinery, New York, NY, USA, 349–352. https://doi.org/10.1145/1216295.1216364

[6] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz,

Brandyn White, Samual White, and Tom Yeh. 2010. *VizWiz: Nearly Real-Time Answers to Visual Questions.* Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/1866029.1866080

[7] Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 12458–12467. https://doi.org/10.1109/CVPR.2019.01275

[8] Luciano Floridi; Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30, 4 (2020), 681–694. https://doi.org/10.1007/s11023-020-09548-1

[9] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.07289

[10] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to Evaluate Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[11] Dhruba Dahal and Raju Shrestha. 2019. Accessible Image Description Using Sample Example Cues. In *Smart Accessibility 2019, Proceedings of the Fourth International Conference on Universal Accessibility in the Internet of Things and Smart Environments.* IARIA, 6–9.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[13] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2021. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. arXiv:2012.11696 [cs.CV]

[14] Himmat Dogra. 2020. *A Framework for an automatic evaluation of image description based on an image accessibility guideline.* Master's thesis. Oslo Metropolitan University (OsloMet).

[15] Hulya Francis, Dhiya Al-Jumeily, and Tom Oliver Lund. 2013. A Framework to Support E-Commerce Development for People with Visual Impairment. In *2013 Sixth International Conference on Developments in eSystems Engineering.* 335–341. https://doi.org/10.1109/DeSE.2013.66

[16] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19).* Association for Computing Machinery, New York, NY, USA, 549–559. https://doi.org/10.1145/3308558.3313605

[17] Ramiro Gonçalves, José Martins, and Frederico Branco. 2014. A Review on the Portuguese Enterprises Web Accessibility Levels – A Website Accessibility High Level Improvement Proposal. *Procedia Computer Science* 27 (2014), 176–185. https://doi.org/10.1016/j.procs.2014.02.021 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, DSAI 2013.

[18] Elisa Kreiss, Noah D. Goodman, and Christopher Potts. 2021. Concadia: Tackling image accessibility with context. arXiv:2104.08376 [cs.CL]

[19] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the ACL-04* (2004), 74–81.

[20] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[21] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. *Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images.* Association for Computing Machinery, New York, NY, USA, 5988–5999. https://doi.org/10.1145/3025453.3025814

[22] Jayawant N. Mandrekar. 2010. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology* 5, 9 (2010), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

[23] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. 2015. Guiding Novice Web Workers in Making Image Descriptions Using Templates. *ACM Trans. Access. Comput.* 7, 4, Article 12 (Nov. 2015), 21 pages. https://doi.org/10.1145/2764916

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* 311–318.

[25] Raju Shrestha. 2021. A Neural Network Model and Framework for an Automatic Evaluation of Image Descriptions based on NCAM Image Accessibility Guidelines. In *The 4th Artificial Intelligence and Cloud Computing Conference (AICCC), Kyoto, Japan.* ACM. (In process of publication).

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* (2017), 5998–6008. arXiv:1706.03762 [cs.CL]

[27] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[28] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2019. A Comparison of Transformer and LSTM Encoder Decoder Models for ASR. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).* 8–15. https://doi.org/10.1109/ASRU46091.2019.9004025