A Neural Network Model and Framework for an Automatic Evaluation of Image Descriptions based on NCAM Image Accessibility Guidelines

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

ABSTRACT

Millions of people who are either blind or visually impaired have difficulty understanding the content in an image. To address the problem textual image descriptions or captions are provided separately or as alternative texts on the web so that the users can read them through a screen reader. However, most of the image descriptions provided are inadequate to make them accessible enough. Image descriptions could be written either manually or automatically generated using software tools. There are tools, methods, and metrics used to evaluate the quality of the generated text. However, almost all of them are word-similarity-based and generic. Even though there are standard guidelines such as WCAG2.0 and NCAM image accessibility guidelines, they are rarely used in the evaluation of image descriptions. In this paper, we propose a neural network-based framework and models for an automatic evaluation of image descriptions in terms of compliance with the NCAM guidelines. A custom dataset was created from a widely used Flickr8K dataset to train and test the models. The experimental results show the proposed framework performing very well with an average accuracy of above 98%. We believe that the framework could be helpful and useful for the authors of image descriptions in writing accessible image descriptions for the users.

CCS CONCEPTS

• **Computing methodologies**; • **Neural networks**;

KEYWORDS

Image accessibility, Image description, Image caption, Automatic evaluation, NCAM guidelines, Machine learning, Neural network

ACM Reference Format:
Raju Shrestha. 2021. A Neural Network Model and Framework for an Automatic Evaluation of Image Descriptions based on NCAM Image Accessibility Guidelines. In *2021 4th Artificial Intelligence and Cloud Computing Conference (AICCC '21), December 17–19, 2021, Kyoto, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3508259.3508269

AICCC '21, December 17–19, 2021, Kyoto, Japan

ACM ISBN 978-1-4503-8416-2/21/12…$15.00
https://doi.org/10.1145/3508259.3508269

1 INTRODUCTION

Tim Berners-Lee, the inventor of the world wide web, states that "The power of the web is in its universality. Access to the contents by everyone regardless of disability is an essential aspect." All types of digital content such as on the web should be accessible to all kinds of people including people with disabilities. Web accessibility aims at overcoming the barriers and making web pages accessible for people with disabilities. This paper is mainly focused on image accessibility; images being an integral part of the digital content including the Internet.

Images can be of several types such as functional images, informative images, decorative images, group of images, and map images. Image description or image caption is commonly used to convey information about the images, which people with visual impairment can read through a screen reader. Images without descriptions become inaccessible to users with visual impairment.

W3C (World Wide Web Consortium) recommends providing image descriptions through alternative text (ALT text) to make them accessible [1]. However currently, most of the images on the Internet are either without any alternative text or containing inaccessible alternative text [2].

An image description should describe the content, context, and purpose of the image [3]. Just providing an image description may not be good enough if it doesn't convey the desired information about the image in an easy and understandable way. In other words, image descriptions should be of good quality for better accessibility. There are accessibility guidelines such as WCAG2.0 (Web Content Accessibility Guidelines, version 2.0) and NCAM (National Center for Accessible Media) which help write accessible image descriptions. However, most of the content on the web currently is not adhering to any accessibility guidelines [3].

Literature suggests that a reason for inaccessible image descriptions could be because of the complexity of writing image descriptions, lack of professional web authors with good knowledge and importance of accessibility and accessibility tools [4], and also due to the lack of time or interest in the guidelines. Moreover, some of the accessibility tools may not support new or changed accessibility guidelines [5]. Therefore, evaluation of the quality of an image description in terms of its compliance with a standard accessibility guideline is important as it would be helpful for the authors of image descriptions.

Evaluation of accessibility of image descriptions can be done either manually, usually by accessibility experts or automatically using a software tool. Dahal and Shrestha [6] have shown a manual evaluation of image descriptions based on NCAM guidelines. Manual evaluation is obviously tedious and time-consuming. Therefore,

there is a need for an automated evaluation method that enables quicker evaluation and ensures the quality of image descriptions in terms of accessibility. Most of the existing automatic evaluations of image descriptions are based on rule-based metrics such as BLEU [7], ROUGE [8], METEOR [9], etc.

This work aims at developing a framework that can automatically evaluate image descriptions based on NCAM image accessibility guidelines. The framework is developed using neural networks, which is one of the most widely used machine learning techniques.

The rest of the paper is organized as follows. Section 2 presents the background and related works. Section 3 describes the proposed framework which includes the NCAM guidelines used, Dataset, neural network model, and performance evaluation metrics. Section 4 describes the experiments and presents and discusses the results. Finally, Section 5 concludes the paper.

2 BACKGROUND AND RELATED WORKS

In general, evaluation of image descriptions can be carried out either manually or automatically. Manual evaluation is carried out by experts or web authors whereas automatic evaluation uses computer algorithms.

Dahal and Shrestha [6] proposed a sample example cue-based image description authoring for improving the accessibility of image descriptions. They found that providing random sample example image(s) with accessible image descriptions helps authors writing image descriptions with better accessibility. Vázquez and Lehmann [10] presented Acrolinx language checker software that can be used as an accessibility evaluation tool. Customized Acrolinx was used to verify the alternative text in the web images, however, it has limitations in customizing its functionalities.

There are works being done that automatically evaluate image descriptions using rule and word similarity-based metrics such as BLEU [7], METEOR [8], ROUGE [9], and CIDer [11]. Hodosh et al. [12] point out limitations and little usefulness of these metrics because of less similarity of these evaluation metrics with the human judgments. They proposed a framework that evaluates image descriptions, which uses Kernel Canonical Correlation Analysis (KCCA) to capture the lexical-based similarity between the words and rank tasks that correlates highly with human judgments.

Bigham [13] proposed a classifier that can evaluate the accessibility of alternative text on web pages. The classifier was trained using labeled examples from a dataset and performed with limited accuracy of about 86%.

None of these works use any accessibility guidelines in their evaluation of image descriptions, justifying a need for one for accessible image descriptions.

3 PROPOSED MODEL AND FRAMEWORK

The proposed framework for the evaluation of an accessible image description includes four parts: NCAM image accessibility guidelines, dataset, a machine learning model, and performance evaluation. They are described in the following sub-sections.

3.1 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 [6], 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.
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.

3.2 Dataset

There are various image captioning datasets available such as MS-COCO, ImageNet, Flickr8K, and Flickr30K. However, none of them are intended for evaluations of their quality based on any standard accessibility guidelines. Therefore, we created a custom dataset from the Flickr8K dataset in Dogra's master thesis [14]. Flickr8K dataset[1] was selected as the size was manageable for manual labeling and also because of the faster training, as suggested by Shinde [15]. It is one of the most used image captioning datasets consisting of 8000 images each paired with five different captions describing the image. A set (4th) among the five sets of descriptions in the Flickr8K dataset was used to create an 8K labeled dataset.

Image captions were labeled manually by experts who have good knowledge about image accessibility and universal design. Based on their understanding of the ten NCAM guidelines and perception of the given image descriptions, the experts gave percentage scores to the image descriptions in terms of their compliance with the selected ten NCAM guidelines.

To include various aspects of NCAM guidelines, some captions have been edited to introduce examples of symbols, punctuations, spelling errors, grammatical errors, past tenses, emotions, interpretation, and analysis. As manual labeling of compliance could be highly subjective and precise values within one percentage precision would be difficult and may not be required most of the time, the dataset was transformed with the compliance labels within 10% precision (i.e., 10%, 20%, . . ., 100%) and used in this work. Figure 1 shows six sample example images from the dataset along with their image descriptions and compliance labels.

3.3 Neural network model

The objective of the machine learning model is to automatically evaluate a given image caption in terms of the percentage compliance to the ten selected NCAM image accessibility guidelines. It is basically a supervised learning problem. We chose the neural

[1] Flickr8K Dataset: https://www.kaggle.com/adityajn105/flickr8k

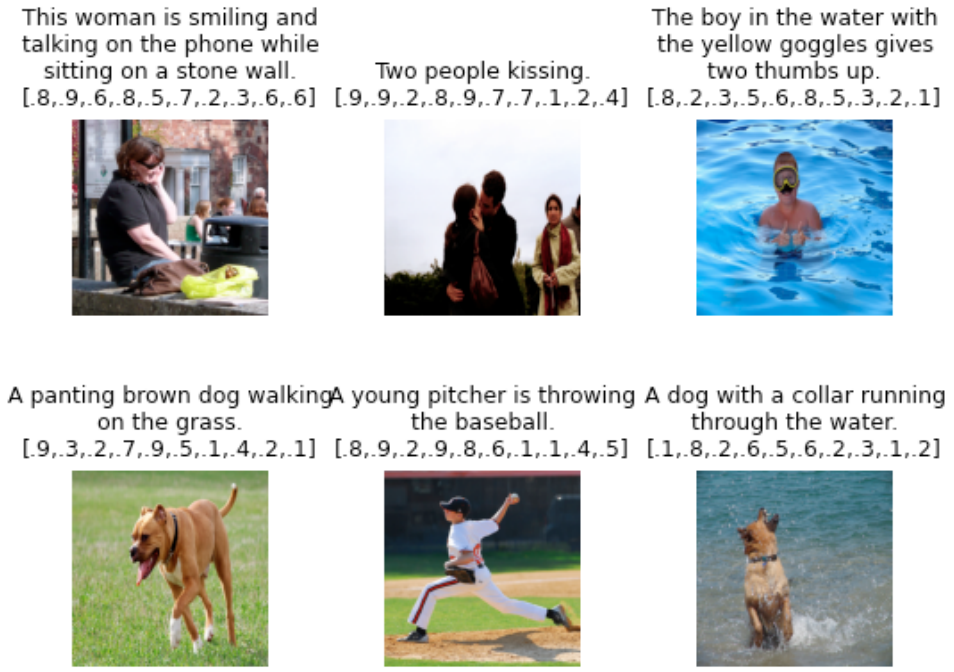

Figure 1: Sample images from the dataset along with the captions and compliance labels to the ten NCM image accessibility guidelines.

network model because of its capability, extensibility, and wider use in the machine learning domain. Ten different neural network models are created corresponding to the ten NCAM guidelines. Feature selection and model architecture are described below.

Feature selection: To train the neural network models, features are selected for the respective model or guideline as follows.

1. Length of the caption, repetition, and similarity of the words therein, and use of qualifiers such as actually, basically, etc. are used to determine succinctness of a given caption. Jaccard similarity matrix [16] is used to calculate similar words.
2. Number of color terms that appeared in the caption and whether the image is a map, chart, or a diagram are used as the features. Webcolors library [17] is used to identify the colors terms.
3. Number of terms such as 'named', 'defined', 'called' etc., words ending with '-ation', and words inside quotes and brackets are used as features to determine new terms and concepts.
4. Readability score, presence of difficult words, and prophanity are used as the features for the understanding and context. Python libraries 'readability' and 'textstat' are used for this.
5. The number of past tense and past participle verbs are used as the features in this case. Part of speech tagging of the Natural Language Toolkit (NLTK) [18] is used for this.
6. Number of spelling errors, grammatical errors, and punctuation errors are used as the features for the model corresponding to guideline #6. Spelling, grammar, and punctuation should be correct. The Python library 'language_tool_python' is used for this.
7. Number of abbreviations such as km, cm, kg, etc., special symbols such as currency symbols and other special characters are used as the features.
8. The model uses the number of words such as image, photo, map, chart, graph, painting, and availability of relevant terms. For example, it is anticipated that the caption containing the term 'map' also contains the terms like 'country', 'region', 'area', and/or the names.
9. Sentiment score from the sentiment analysis of the caption is used as the feature in the model. SentimentIntensityAnalyzer from the Python package 'vaderSentiment' is used to compute sentiment scores.
10. Number of occurrences of the definition words like 'think', 'seems', 'perhaps' etc., modal words like 'may', 'can', and race terms that describes color, religion, gender, ethnicity are used as the features to identify personal interpretation, opinions, and analysis in the given caption.

Model architecture: Figure 2 shows a general neural network architecture, which is made up of L number of hidden layers, each with N number of neurons or units. L2-regularization parameter λ is used to address overfitting. Adam (Adaptive moment) [19] optimizer, which is arguably the most popular and effective optimizer in deep learning, is used with default learning rate in TensorFlow 2.6. To capture non-linear relationships, Leaky ReLU activation is used in the hidden layers. Mean square error is used as the loss function in the neural network models. Since the number of features is small, polynomial features with degree, d, are used as it not only generates additional features but is also found to capture relationships better. It has been found that the polynomial features improve model

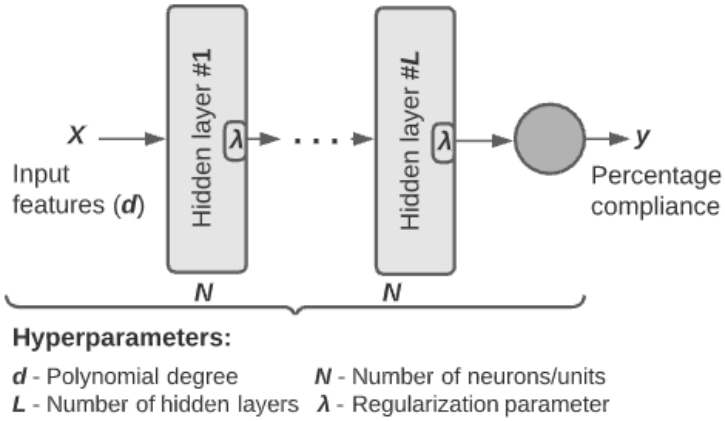

Figure 2: A general neural network architecture used to create the ten models corresponding to the ten NCAM guidelines.

performance. Ten specific models corresponding to the ten NCM guidelines are obtained from the general model through hyperparameter optimization using the four hyperparameters, d, L, N, and λ.

The models were trained, tested, and evaluated using the dataset (see Section 3.2). Section 4 below describes the training and testing process. The next sub-section presents the metrics used for the performance evaluation of the models.

3.4 Performance evaluation

Two metrics are used to evaluate the performance of the models in predicting compliance level to the NCAM guidelines: prediction error, and accuracy. These metrics are defined below.

Prediction error (error in short) is defined as the absolute difference between the predicted value and the target value.

$$error = |predicted\ value - target\ value|$$

Accuracy is defined as the ratio of correct prediction to the total number of predictions:

$$accuracy = \frac{correct\ predictions}{total\ number\ of\ predictions} \times 100\%$$

A prediction is considered correct if the predicted value is within 1 percent of the target value, i.e., if the error is less than 0.01. 1 percent is reasonable & acceptable for the 10 percent precision of a compliance label.

4 EXPERIMENTS, RESULTS, AND DISCUSSION

The dataset created as described in Section 3.2 was shuffled and split into training, validation, and test sets in the ratio of 70:15:15 percent. Using the training and validation data, the proposed neural network model (see Section 3.3) was tuned for the ten models corresponding to the ten NCAM guidelines through hyperparameter optimization (HPO) process in RayTune[2] using Bayesian optimization technique with minimum validation error as the optimization metric. A fixed batch size of 32 was used in all ten models.

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

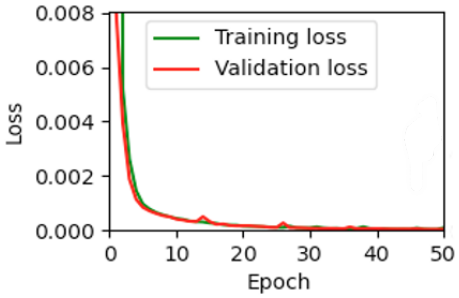

Figure 3: Learning curves of model #1.

Table 1 shows the optimal hyperparameter values for the ten models obtained from the HPO process. As an illustration, Figure 3 shows learning curves from the training of model #1. Other models also produce similar learning curves and all the models converged well in 50 epochs.

The optimized hyperparameter values in Table 1 show that the models are simple neural networks with one or two hidden layers, with the number of units ranging from 16 to 2048. The degree of polynomial feature ranges from 1 to 7. Combinations of L2-regularization and dropout with their respective values shown in Table 1 helped mitigate overfitting in the corresponding models.

Next, the models were tested with the test data and evaluated their performance using the error and accuracy metrics described in Section 3.4. The resulting metric values are given in Table 2. Figure 4 shows the results in graphical plots. The results show a very good performance from the models with an average error 0.004 and an average accuracy of 98% from all ten guidelines. Individual model-wise, the lowest accuracy is about 94% from model #6. In terms of errors, model #9 produced relatively higher error of around 0.012, which is still reasonably low.

As we know that the selection of features is the most challenging task in a machine learning modeling. This is more true here due to the difficulty in interpreting the guidelines consistently. However, we see that the models produced very good results with the selected

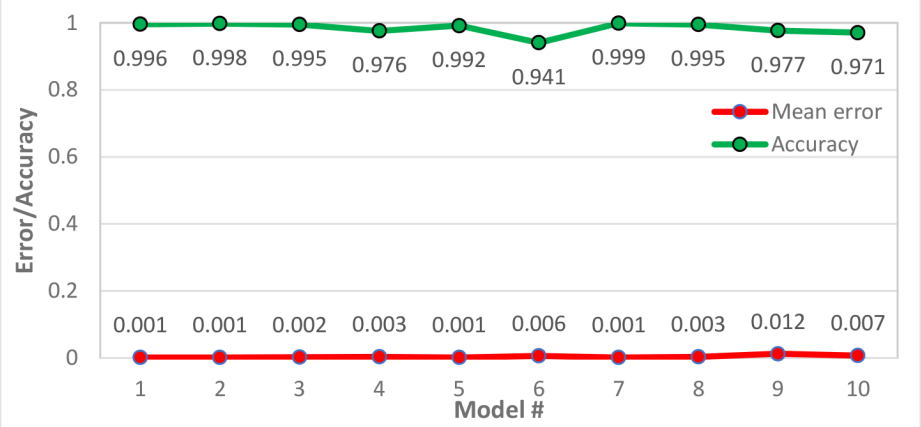

Figure 4: Plots showing accuracies and prediction errors along with standard errors for the models corresponding to the ten guidelines.

Table 1: Optimized hyperparameter values for the ten models.

Model #	*d*	*L*	*N*	λ
1	5	1	16	0.00010
2	5	1	64	0.00010
3	3	1	2048	0.00019
4	3	2	512	0.00011
5	2	1	1024	0.00011
6	7	2	128	0.00010
7	3	1	1024	0.00010
8	1	1	2048	0.00012
9	5	1	1024	0.00019
10	3	1	512	0.00012

Table 2: Mean error and accuracy from the models.

Model #	Mean error	Accuracy
1	0.001	0.995
2	0.001	0.998
3	0.002	0.995
4	0.003	0.976
5	0.001	0.992
6	0.006	0.941
7	0.001	0.999
8	0.003	0.995
9	0.012	0.977
10	0.007	0.971
Average	**0.004**	**0.984**

small number of manually extracted features. The performance could possibly be improved further with more extended features.

It is to be noted that this work assumes that the given image description well describes the image both in terms of its content and context and the sole purpose here is to evaluate it in terms of image accessibility. A potential future work would be to incorporate the quality of the content and context from the image in the evaluation.

5 CONCLUSION

The proposed framework based on neural network machine learning models for the evaluation of image captions is novel in the sense that it evaluates the quality of image descriptions in terms of the widely used standard image accessibility guidelines, NCAM. The models performed very well with an average of above 98% accuracy in the dataset created using the Flickr8K dataset.

It is believed that the framework could be helpful for web authors and image describers to provide high-quality image descriptions or captions to the users for better image accessibility.

REFERENCES

[1] W3C. *HTML5: Techniques for providing useful text alternatives.* 2016. https://w3c.github.io/alt-techniques/.
[2] P. Bigham, J., *et al.*, *WebInSight: Making web images accessible.* Vol. 2006. 2006. 181-188.
[3] Petrie, H., C. Harrison, and S. Dev, *Describing images on the web: A survey of current practice and prospects for the future.*Proceedings of Human Computer Interaction International (HCII), 2005.**71**.
[4] Abuaddous, H.Y., M.Z. Jali, and N. Basir, *Web accessibility challenges.*International Journal of Advanced Computer Science and Applications, 2016. **7** (10): p. 172-181.
[5] Trewin, S., *et al. Accessibility challenges and tool features: An IBM Web developer perspective.* in *Proceedings of the 2010 international cross disciplinary conference on web accessibility (W4A).* 2010. ACM.
[6] Dahal, D. and R. Shrestha, *Accessible Image Description Using Sample Example Cues,* in *The Fourth International Conference on Universal Accessibility in the Internet of Things and Smart Environments.* 2019: Athens, Greece. p. 13-16.
[7] Papineni, K., *et al. BLEU: A method for automatic evaluation of machine translation.* in *Proceedings of the 40th annual meeting on association for computational linguistics.* 2002. Association for Computational Linguistics.
[8] Banerjee, S. and A. Lavie. *METEOR: An Automatic Metric for MT Evaluation with-Improved Correlation with Human Judgments.* in *The ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translationand/or Summarization.* 2005. Association for Computational Linguistics (ACL).
[9] Lin, C.-Y. *ROUGE: A package for automatic evaluation of summaries.* in *Text summarization branches out.* 2004.
[10] Vázquez, S.R. and S. Lehmann. *Acrolinx: A controlled-language checker turned into an accessibility evaluation tool for image text alternatives.* in *Proceedings of the 12th Web for All Conference.* 2015. ACM.

[11] Vedantam, R., C.L. Zitnick, and D. Parikh. *CIDEr: Consensus-Based Image Description Evaluation.* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015. IEEE.

[12] Hodosh, M., P. Young, and J. Hockenmaier, *Framing image description as a ranking task: Data, models and evaluation metrics.*Journal of Artificial Intelligence Research, 2013.**47**: p. 853-899.

[13] Bigham, J.P. *Increasing web accessibility by automatically judging alternative text quality.* in *Proceedings of the 12th international conference on Intelligent user interfaces.* 2007. ACM.

[14] Dogra, H.K., *A Framework for an automatic evaluation of image description based on an image accessibility guideline, Master thesis at Department of Computer Science.* 2020, Oslo Metropolitan University (OsloMet).

[15] Shinde, R. *Image Captioning With Flickr8k Dataset & BLEU.* 2019. https://medium.com/@raman.shinde15/image-captioning-with-flickr8k-dataset-bleu-4bcba0b52926.

[16] Sieg, A. *Text Similarities: Estimate the degree of similarity between two texts.* 2018. https://medium.com/@adriensieg/text-similarities-da019229c894.

[17] Bennett, J. *Webcolors: Module contents.* 2014. https://webcolors.readthedocs.io/en/1.5/contents.html.

[18] Rachiele, G. *Tokenization and Parts of Speech (POS) Tagging in Python's NLTK library.* 2018. https://medium.com/@gianpaul.r/tokenization-and-parts-of-speech-pos-tagging-in-pythons-nltk-library-2d30f70af13b.

[19] Kingama, D. and J. Ba. *Adam: A method for stochastic optimization.* in *The 3rd International Conference for Learning Representations (ICLR).* 2015. San Diego, USA