Technical Disclosure Commons

Defensive Publications Series

December 19, 2018

Machine-learned alt text for images and videos

Sandro Feuz

Mohammadamin Barekatain

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation

Feuz, Sandro and Barekatain, Mohammadamin, "Machine-learned alt text for images and videos", Technical Disclosure Commons, (December 19, 2018)
https://www.tdcommons.org/dpubs_series/1791

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

Machine-learned alt text for images and videos

ABSTRACT

HTML has a provision for adding a textual description, known as alt text, to media elements in a webpage, such as images and videos. This is useful for improving accessibility of web pages, for display environments that do not support a particular media item, for content searching, etc. The alt text is currently added manually, e.g., by the web page developer or publisher. This is a labor intensive activity and does not work for dynamic media. This disclosure utilizes machine learning techniques to automatically add alt text to media content.

KEYWORDS

- alt text
- alt attribute
- webpage
- accessibility
- media content
- scene classification
- video summarization
- object classification

BACKGROUND

HTML has a provision for adding a textual description, known as alt text, to media elements in a webpage, such as images and videos. This is useful for improving accessibility of web pages, for display environments that do not support a particular media item, for content searching, for search-engine optimization, etc. The alt text is currently added manually, e.g., by

the web page developer or publisher. This is a labor intensive activity and does not work for dynamic media.

DESCRIPTION

Per the techniques of this disclosure, a generative machine learning model, e.g., as described in [1], is utilized to generate a summary of an image, video, or other media item. The model is integrated into a web browser and automatically generates alt text for media that is loaded without publisher-supplied alt text, e.g., in an alt attribute in the HTML source for a webpage. The user is provided with an option to turn off automatic alt text generation.

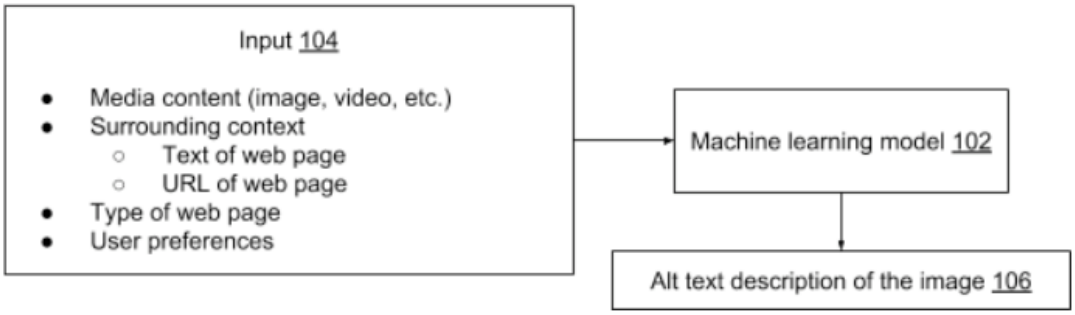

Fig. 1: Machine-learned generation of alt text

Fig. 1 illustrates generation of alt text using a machine learning model, per techniques of this disclosure. A machine learning model (102) takes as input (104) the following features:

- media content to be described using alt text, e.g., image, video, etc.;
- context surrounding the media, e.g., text, web page URL, etc. (e.g., that can be utilized to generate a more contextualized description of the media item);
- type of web page (e.g., news, sports, etc.);
- user preferences and user-specific features, obtained with user permission; etc.

The model is trained by supplying pairs of media items and the corresponding manually-labeled alt text. Existing web pages can be utilized as sources of such training data. Videos are summarized by supplying frames from the video to the generative machine learning model.

The machine learning model can be, e.g., a generative machine learning model. Example ways to implement a generative machine learning model include Variational Auto Encoders (VAEs) or Generative Adversarial Networks (GANs). Each of these types of models can include long short-term memory (LSTM) neural networks, recurrent neural networks, convolutional neural networks, etc. Other machine learning models, e.g., support vector machines, random forests, boosted decision trees, etc., can also be used.

Example: The features supplied to the machine learning model indicate a user interest in movies. Alt text for movie-related images generated by the machine-learning model can include text such as “Men at an action movie set,” or specifically include the name of the movie, names of actors in the scene or the movie, or others associated with the movie etc.

Multiple language support

Alt text can be provided in multiple languages for alt-text, for example, by training separate models for each language and adding the models to the language packs of the browser. Alternatively, a generative model can be trained to generate alt text in one language which is then machine-translated to other languages.

Resource optimization

To avoid the inefficiencies of re-running inferences on the same media item, the generated alt-text for a given media item can be cached, e.g., at a server, and indexed, e.g., using a checksum of the media item. If the user permits, the server is queried to check if alt text for the

item is available on the server prior to running alt text generation on a media item. If such text is found on the server, it is used for the particular media item. Otherwise, the alt text is obtained using the aforementioned techniques and, if the user permits, is sent to the server for storage and future retrieval. This technique of storing pre-generated alt text also provides for a mix of curated and auto-generated alt text.

In this manner, the techniques for auto-generation of alt text described herein reduce the need for manual input by the website developers and publishers. The techniques also work for dynamic media, e.g., images or videos that change over time. The generated alt text is tailored to the user.

The techniques of automatic alt text generation can be implemented as part of a web browser application, or any other software application that displays online content. For example, such applications can include mobile and desktop apps for image viewing and sharing, social networking, news/magazine apps, etc.

In situations in which certain implementations discussed herein may collect or use personal information about users (e.g., user data, information about a user's social network, user's location and time at the location, user's biometric information, user's activities and demographic information), users are provided with one or more opportunities to control whether information is collected, whether the personal information is stored, whether the personal information is used, and how the information is collected about the user, stored and used. That is, the systems and methods discussed herein collect, store and/or use user personal information specifically upon receiving explicit authorization from the relevant users to do so.

For example, a user is provided with control over whether programs or features collect user information about that particular user or other users relevant to the program or feature. Each

user for which personal information is to be collected is presented with one or more options to allow control over the information collection relevant to that user, to provide permission or authorization as to whether the information is collected and as to which portions of the information are to be collected. For example, users can be provided with one or more such control options over a communication network. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. As one example, a user's identity may be treated so that no personally identifiable information can be determined. As another example, a user's geographic location may be generalized to a larger region so that the user's particular location cannot be determined.

CONCLUSION

The techniques of this disclosure use machine learning to automatically add alt text to media content such as images, video, and other content on webpages and other online content.

REFERENCES

[1] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156-3164. 2015. https://arxiv.org/abs/1411.4555 accessed on Nov. 18, 2018.