Automatic Context-Aware Image Captioning

Sandro Feuz
Sebastian Millius

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

Recommended Citation

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.
AUTOMATIC CONTEXT-AWARE IMAGE CAPTIONING

ABSTRACT

A system and method are disclosed that automatically trigger system-generated captions for images. The system includes a machine learning model and an interface that proposes captions for an image that a user intends to share. The machine learning model is generated by considering the factors such as: extraction of metadata from the image, semantic information extracted from raw image, context of the current chat, image captions from previously shared images, personal context or user modifications to the image. Based on the results of the machine learning model, the system generates and proposes several captions to the user. The user then selects a suitable caption and shares the image. The user may also modify or manually enter the caption. The system generated image captioning enables easy sharing of captioned images which are personalized and artistic.

BACKGROUND

Generally, images constitute a large fraction of chat messages. The images are often sent along with a message (caption) describing it. For instance, a user sends an image from the honeymoon with his wife to a friend, with the message: “Alice and I enjoying our honeymoon in the Bahamas”. Currently, captions for images are manually inserted in all available chat applications.

DESCRIPTION

A system and method are disclosed that triggers system-generated captions for images. The system includes a machine learning model and an interface that proposes captions for an image that a user intends to share. The method includes a data-driven, trainable machine learning approach that determines the chat context for the images and generates captions when sending
images in chat applications as shown in FIG. 1. When the user begins to share an image, the machine learning model acquires inputs for generating suitable captions. The inputs may include image metadata, image content, chat context and the personalized chat history. The system then generates the captions based on the available information and presents it to the user as illustrated in FIG. 2. The user selects the desired caption and sends the captioned image.

FIG. 1: Method to automatically generate context-aware captions for images
The machine learning model may operate in a way similar to existing image description generating models. Training the machine-learning model to create personalized captions involves taking various features as inputs and generating textual representation of captions, or language independent intermediate representation to create a textual representation out of such captions.

Various features may serve as inputs for the machine learning model. Attributes deduced from the image and image metadata may be used, for instance, the raw image (pixel values), the
image metadata (when (timestamp) and where (coordinates) was the image taken, who took it). For example, geocoded address may be used (“Pelican Bay Hotel - Seahorse Road Bahamas), as can a resolved timestamp (for example, “Super Bowl”, or “First day of School”). Semantic information may be extracted from a raw image, as a result of image content analysis (e.g. Eiffel Tower). Context of the current chat may be used, and may involve analyzing history for the current chat. Image captions from previously shared images may also be a factor. Personal context such as calendar entries (“Honeymoon”), social contacts with their profile pictures, etc. may be used. User modifications such as applied filters, added text - e.g. style-specific proposals may be used. The system then outputs a textual representation of possible captions. Alternatively, the system could create a language-independent intermediate representation to create a textual representation out of, prior to outputting the textual representation.

The second part automatically proposes the generated captions before the user sends an image. Whenever the user selects an image in a chat application and attempts to enter a caption (or accompanying message) to the image, the system extracts the necessary features from the chat history and the image the user selected, and runs them through the machine learning model, to propose captions. The system generates multiple proposals for captions which may be presented to the user as illustrated in FIG. 1, and the user may then select one of them without manually typing it. Optionally, the user may modify the caption before sending.

The machine learning model may be trained in a supervised fashion by providing the input and desired output data. Training data may be accumulated by analyzing chat logs and extracting existing captions, which were entered manually by the users. The model may learn natural captions from the way human users annotate images they share and also from the chat history. The machine learning model may also use a deep recurrent neural network (RNN) with
additional convolutional layers to extract semantic information from raw images.

Also, the system may include a rule-based approach for default image captioning. This may be used to combine extracted labels/content from an image with manually curated caption templates. For example, the system would detect that the image was taken in Paris, and user may have a list of templates, for example, “A day in Paris”, “Having fun in Paris”. Furthermore, the technology may be extended to provide captions of video content or animated GIFs.

For example, as shown in FIG. 2, the user selects an image with a view over a city from a restaurant. The extracted features may include the semantic location deduced from the exact location, which in this case is “Casino XY, Las Vegas”, the calendar entries from the user overlapping with the time the photo was taken (Stay at Casino XY, Stag night). Together with the raw image and features extracted from the image content analysis system (keywords like “Gambling”, “Poker”, “Casino”), are fed into the machine learning model. The machine learning model produces multiple example sentences. The top 3 candidates are used as quick-select options, and the user is still left with an option to manually enter the caption. As the system takes into account the user’s modifications to the image and adapts the proposed captions, more “artistic” proposals may be generated. In another example, the usage of chat history and context in which an image is shared makes the proposals more natural and self-referential to the chat history “A fun night with Peter and Sarah at Olive Garden” may be suggested when the friends Peter and Sarah were mentioned in the chat.

The disclosed system and method for image captioning enables simple and easier sharing of images. The scope, flexibility and predictive performance of a machine learned system goes far beyond simple primitive template proposals, giving the user a wide choice of proposals and increasing the potential of usability.