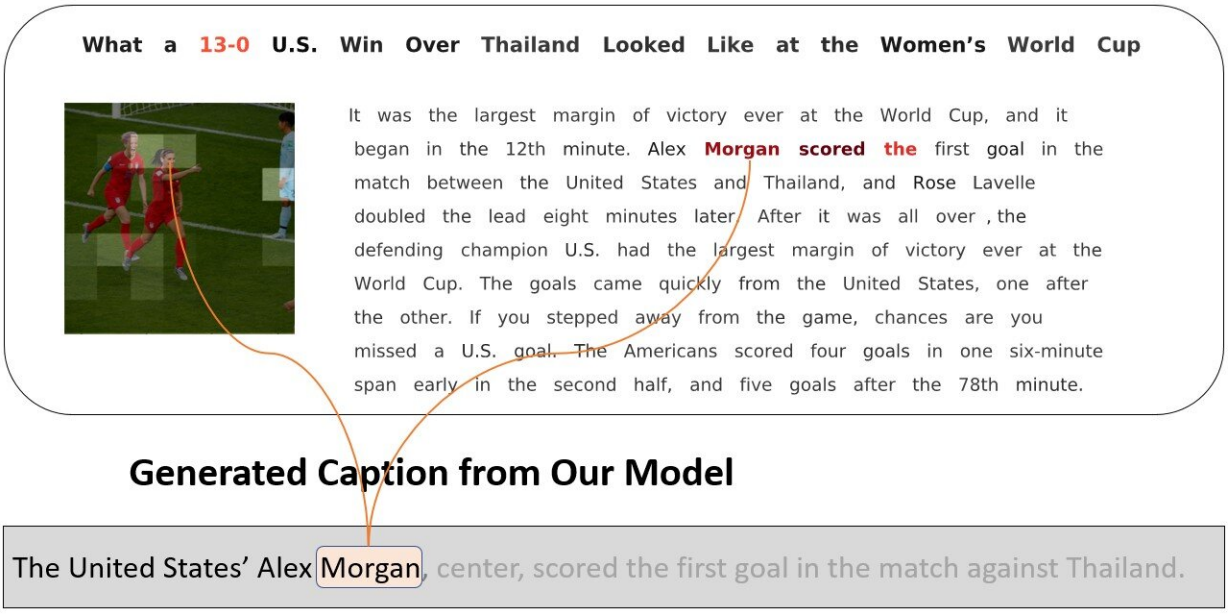
A system to produce context-aware captions for news images

May 18 2020, by Ingrid Fadelli



Given a news article and an image (top), the researchers' model generates a relevant caption (bottom) by attending to the context associated with the image. The attention scores over the image patches and the article text are shown as the decoder generates the word 'Morgan'. Image patches with higher attention have a lighter shade, while highly attended to words are in red. The orange lines point to the highly attended regions. Credit: Tran, Mathews & Xie.

Computer systems that can automatically generate image captions have been around for several years. While many of these techniques perform considerably well, the captions they produce are typically generic and

somewhat uninteresting, containing simple descriptions such as "a dog is barking" or "a man is sitting on a bench."

Alasdair Tran, Alexander Mathews and Lexing Xie at the Australian National University have been trying to develop new systems that can generate more sophisticated and descriptive image captions. In a paper recently pre-published on arXiv, they introduced an automatic captioning system for news images that takes the general context behind an image into account while generating new captions. The goal of their study was to enable the creation of captions that are more detailed and more closely resemble those written by humans.

"We want to go beyond merely describing the obvious and boring visual details of an image," Xie told TechXplore. "Our lab has already done work that makes image captions sentimental and romantic, and this work is a continuation on a different dimension. In this new direction, we wanted to focus on the context."

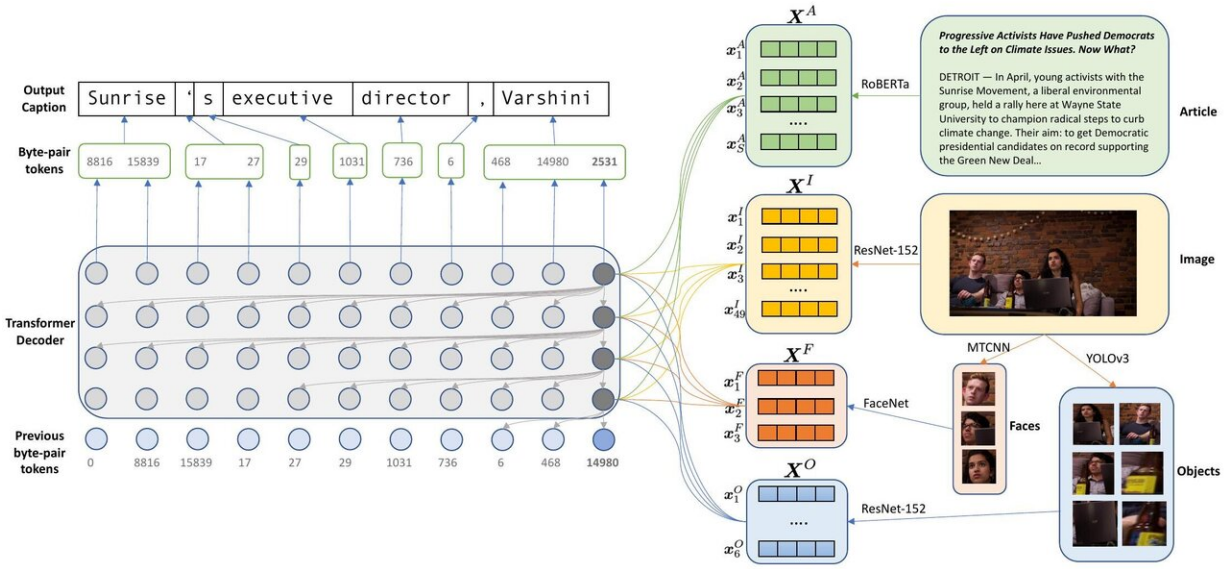
In real-life scenarios, most images come with a personal, unique story. An image of a child, for instance, might have be taken at a birthday party or during a family picnic.

Images published in a newspaper or on an online media site are typically accompanied by an article that provides further information about the specific event or person captured in them. Most existing systems for generating image captions do not consider this information and treat an image as an isolated object, completely disregarding the text accompanying it.

"We asked ourselves the following question: Given a news article and an image, can we build a model that could be aware of both the image and the article text in order to generate a caption with interesting information that cannot simply be inferred from looking at the image alone?" Tran

said.

The three researchers went on to develop and implement the first end-to-end system that can generate captions for news images. The main advantage of end-to-end models is their simplicity. This simplicity ultimately allows the researchers' model to be linguistically rich and generate real-world knowledge such as the names of people and places.



Model overview. Left: Decoder with four transformer blocks; Right: Encoder for article, image, faces, and objects. The decoder takes byte-pair tokens (blue circles at the bottom) as input embeddings. For example, the input in the final time step, 14980, represents ``arsh'' in ``Varshini'') from the previous time step. The grey arrows show the convolutions in the final time step in each block. Colored arrows show attention to the four domains on the right: article text (green lines), image patches (yellow lines), faces (orange lines), and objects (blue lines). The final decoder outputs are byte-pair tokens, which are then combined to form whole words and punctuations. Credit: Tran, Mathews & Xie.

"Previous state-of-the-art news captioning systems had a limited vocabulary size, and in order to generate rare names, they had to go through two distinct stages: generating a template such as "PERSON is standing in LOCATION"; and then filling in the placeholders with actual names in the text," Tran said. "We wanted to skip this middle step of template generation, so we used a technique called byte pair encoding, in which a word is broken down into many frequently occurring subparts such as 'tion' and 'ing.'"

In contrast with previously developed image captioning systems, the model devised by Tran, Mathews and Xie does not ignore rare words in a text, but instead breaks them apart and analyzes them. This later allows it to generate captions containing an unrestricted vocabulary based on about 50,000 subwords.

"We also observed that in previous works, the captions tended to use simple language, as if it were written by a school student instead of a professional journalist," Tran explained. "We found that this was partly due to the use of a specific model architecture known as LSTM (long short term memory)."

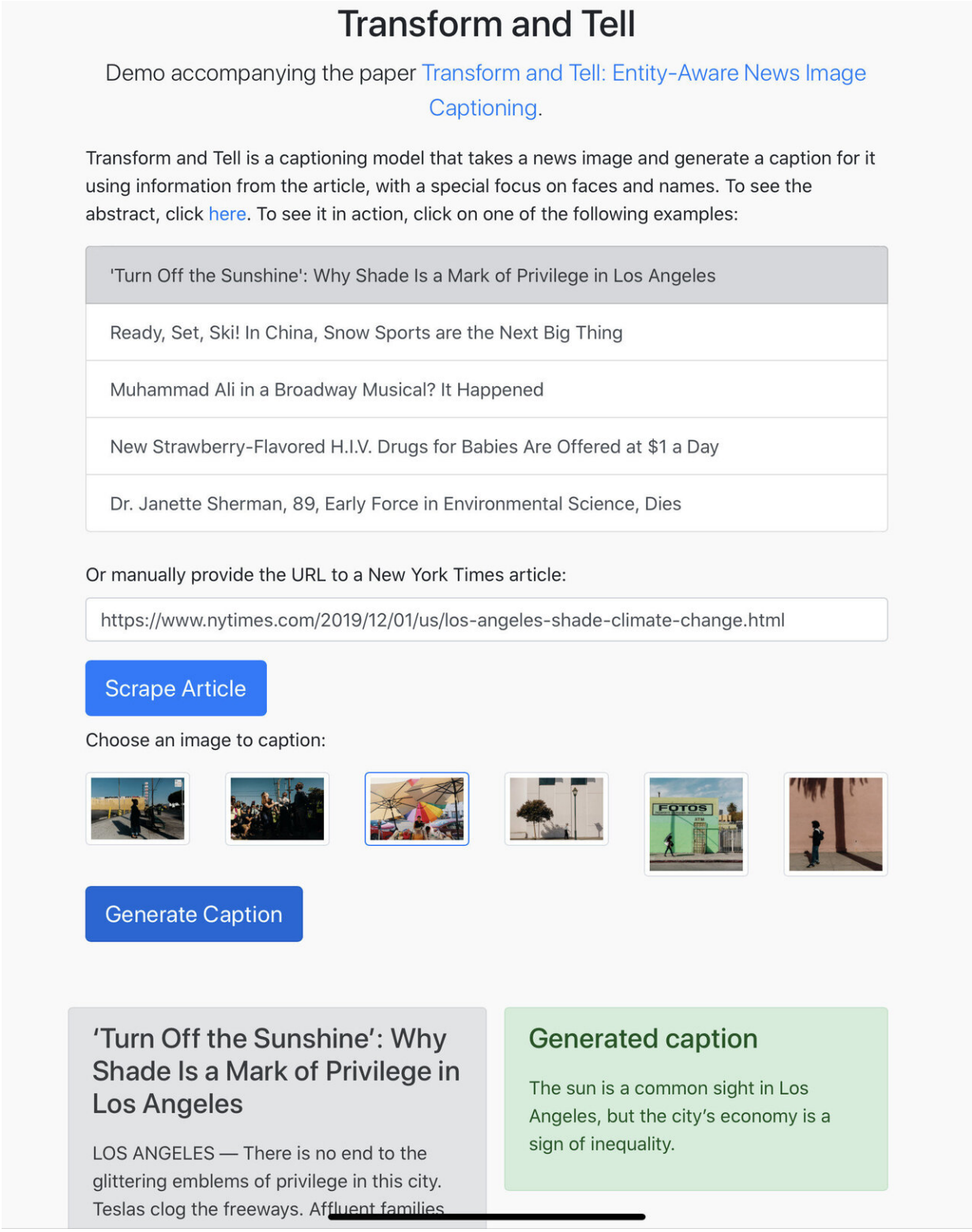
LTSM architectures have become widely used in recent years, particularly to model number or word sequences. However, these models do not always perform well, as they tend to forget the beginning of very long sequences and can take a long time to train.

To overcome these limitations, the research community in language modeling and machine translation has recently started adopting a new type of architecture, dubbed transformer, with highly promising results. Impressed by how these models performed in previous studies, Tran, Mathews and Xie decided to adapt one of them to the image captioning task. Remarkably, they found that captions generated by their transformer architecture were far richer in language than those produced

by LSTM models.

"One key algorithmic component that enables this leap in natural language ability is the attention mechanism, which explicitly computes similarities between any word in the caption and any part of the image context (which can be the article text, the image patches, or faces and objects in the image)," Xie said. "This is done using functions that generalize the vector inner products."

Interestingly, the researchers observed that the majority of images published in newspapers feature people. When they analyzed images published in the *New York Times*, for instance, they found that three-quarters of them contained at least one face.



Screenshot of the captioning system's demo app, which can be accessed at https://transform-and-tell.ml/. Credit: Tran, Mathews & Xie.

Based on this observation, Tran, Mathews and Xie decided to add two extra modules to their model: one specialized in detecting faces and the other in detecting objects. These two modules were found to improve the accuracy with which their model could identify the names of people in images and report them in the captions it produced.

"Getting a machine to think like humans has always been an important goal of artificial intelligence research," Tran said. "We were able to get one step closer to this goal by building a model that can incorporate real-world knowledge about names in existing text."

In initial evaluations, the image captioning system achieved remarkable results, as it was could analyze long texts and identify the most salient parts, generating captions accordingly. Moreover, the captions generated by the model were typically aligned with the writing style of the *New York Times*, which was the key source of its training data.

A demo of this captioning system, dubbed "Transform and Tell," is already available online. In the future, if the full version is shared with the public, it could allow journalists and other media specialists to create captions for news images faster and more efficiently.

"The model that we have so far can only attend to the current article," Tran said. "However, when we look at a news article, we can easily connect the people and events mentioned in the text to other people and events that we have read about the past. One possible direction for future research would be to give the model the ability to also attend to other similar articles, or to a background knowledge source such as Wikipedia. This will give the model a richer context, allowing it to generate more interesting captions."

In their future studies, Tran, Mathews and Xie would also like to train their model to complete a slightly different task to that tackled in their recent work, namely, that of picking an image that could go well with an article from a large database, based on the article text. Their model's attention mechanism could also allow it to identify the best place for the image within the text, which could ultimately speed up news publishing processes.

"Another possible research direction would be to take the transformer architecture that we already have and apply it to a different domain such as writing longer passages of text or summarizing related background knowledge," Xie said. "The summarization task is particularly important in the current age due to the vast amount of data being generated every day. One fun application would be to have the model analyze new *arXiv* papers and suggest interesting content for scientific news releases like this article being written."

More information: Transform and tell: entity-aware news image captioning. arXiv: 2004.08070 [cs.CV]. arxiv.org/abs/2004.08070