Transforming Entities for News Image Captions

Anonymous CVPR submission

Paper ID 7245

Abstract

We propose an end-to-end model to generate image captions in news articles. By combining the transformer architecture, byte-pair encoding, copying with multi-headed attention, and pretrained embeddings from three different modalities (RoBERTa for text, ResNet-152 for images, and FaceNet for faces), our system is able to describe an image with specific named entities mentioned in the article. Our model achieves a CIDEr score of 61 on the GoodNews dataset, significantly outperforming the previous state-of-art CIDEr of 13. We also introduce the NYTimes800k dataset, the largest news image captioning dataset to date. NYTimes800k is an extended version of GoodNews with higher-quality articles and metadata that allow us to study the importance of the image location within the text. On NYTimes800k, we achieve a CIDEr of 65. Pretrained models and source code are available from https://github.com/anonymized-link.

1. Introduction

The internet is home to a huge number of images, many of which lack useful captions. A growing body of work seeks to automatically generate captions that describe the objects and relationships using only visual cues extracted from the image itself [9, 45, 11, 17, 35, 27, 1, 5]. While generic image descriptions have their uses, such as for individuals with vision impairments, they are often of less benefit to the average user. To produce more useful image captions we need to go beyond generic descriptions and introduce information that cannot be gleaned directly from the image alone. Fortunately, many images have an associated context such as a news article, web page, or social media post, which give the image greater meaning than can be extracted from its pixels. To generate captions that go beyond generic description and actually add information that could not be gleaned from the image alone we must take this context into account. We focus on the news image captioning task in order to design practical methods for exploiting contextual information.

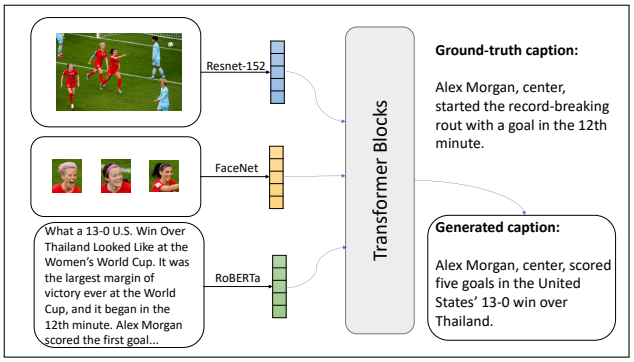

Figure 1: Our transformer model attends to embeddings from three different domains (image patches, faces, and article text). Using byte-pair encoding, the model can then directly produce a caption containing specific named entities without the use of templates.

News image captioning is an interesting instance of contextual captioning where news articles provide context to images.

Captions for news images, such as the example in Figure 1, typically contain details which cannot be derived from the image alone. They also frequently contain proper nouns such as names of people, places, and organizations – in many cases these proper nouns are rare (most people and places do not have many news articles written about them). A system capable of generating high quality news captions should therefor make extensive use of the provided context and be tuned for generating rare proper nouns. Existing approaches to news image captioning [41, 34, 2] rely on text extraction or template filling to deal with rare contextual terms such as names of people and organizations. This makes them relatively inflexibility and means they cannot be trained end-to-end. Moreover, existing approaches do not include specialized visual models for frequent nouns – experiments on the MSCOCO dataset have shown that pre-trained object detectors tuned for frequent nouns lead to more accurate captions [].

This motivates our novel fully end-to-end model for

news image captioning that 1) combines specialised modules for incorporating and selectively attending to image features, human faces, and news article text and 2) applies a state-of-the-art sequence generation model which is able to generate rare tokens, such as proper names, even when they do not form part of the training data. Our model relies on a novel combination of sequence-to-sequence architectures, language representation learning, and vision systems.

In this paper we carefully consider the news image captioning problem and select a set of modeling tools which we combine into a novel architecture that sets a new state-of-the-art result. Our main contributions are threefold:

1. We introduce NYTimes800k, the largest news image captioning dataset to date, containing 446K articles and 794K images with captions from The New York Times spanning 14 years. NYTimes800k builds on the GoodNews dataset; but we write a custom parser to collect higher-quality articles and metadata such as the location of an image within the page.

2. We build a captioning model that combines the power of transformers, byte-pair encoding, copying via multi-headed attention, and attention over three different modalities (text, images, and faces). We show that our model achieves state-of-the-art results with a significant margin over previous methods, and in particular, it can generate names not seen during training without the use of templates.

3. We provide a detailed model analysis, deconstructing the most important modeling components and quantifying the incremental contribution that each of them makes not only to the usual metrics such as BLEU, ROUGE, METEOR, and CIDEr; but also to other linguistic measures like readability scores, caption length, and recall of rare names.

2. Related Works

A large number of methods exist for generating generic image captions that describe objects and relationships using only the image as input. Many of these captioning systems use some combination of a Convolutional Neural Network encoder and an RNN with a closed vocabulary as a decoder [17, 9, 45]. Attention over image patches was introduced in "Show, Attend and Tell" [48], in which the attention weights are obtained by feeding the image embeddings and the previous hidden state of the RNN through a multilayer perception. Many extensions to these models have been proposed such as giving the model the option to not attend to any image region [27], using reinforcement learning to directly optimise for the CIDEr metric [35, 13], and using a bottom-up approach to propose a region to attend to [1]. All of these systems generate restricted vocabulary generic captions without considering context external to the image.

A related task which does consider image context is news image captioning, where the image caption is generated using the article text as context. One key challenge of news image captioning is generating rare entity names, for example the names of people who do not make many media appearances. Early non-neural approaches include extractive methods that use n-gram models to combine existing phrases [12] or simply retrieving the most representative sentence [41] in the article. The neural network approach taken by Ramisa *et al*. [34] was able to generate entirely new text with an LSTM decoder that took as input a word2vec representation of the article concatenated with a CNN representation of the image. Even so this approach was unable produce names that were not seen during training.

To overcome the limitation of a fixed-size vocabulary caption templates can be used. This involves first generating a template sentence with placeholders for named entities, e.g. "PERSON speaks at BUILDING in DATE.". This template can be generated using an LSTM or other sequence generation model [2]. Afterwards, a selection algorithm picks the best candidate for each placeholder. For example, Lu *et al*. [26] built a knowledge graph for each combination of entities and select the most likely combination. Meanwhile, Biten *et al*. [2] filled the template by extracting entities from the sentence in the article which had the highest cosine similarity to the template. One key difference between our proposed model and that of previous approaches[2, 26] is that our model can generate a captions with named entities directly – without using an intermediate template.

One tool that has had seen recent successes in many natural language processing tasks is the transformer neural network. Transformers have been show to consistently outperforming Recurrent Neural Network architectures in language modeling [33], story generation [10], summarization [40], and machine translation [3]. Furthermore, transformer based models such as BERT [8], XLM [20], XLNet [49], RoBERTa [24], and ALBERT [21] have been shown to produce high level text representations suitable for transfer learning. Furthermore, using byte-pair encoding (BPE) [38] to represent uncommon words as a sequence of subword units can enable the transformer function in an open vocabulary setting.

Transformers have been shown to yield competitive results in generating generic MS COCO captions [53, 22]. Zhao *et al*. [52] have gone further and trained transformers to produce some named entities in the Conceptual Captions dataset [39]. However, this dataset provides no additional context to the image, making it a different problem to news image captioning. Moreover, the authors used web-

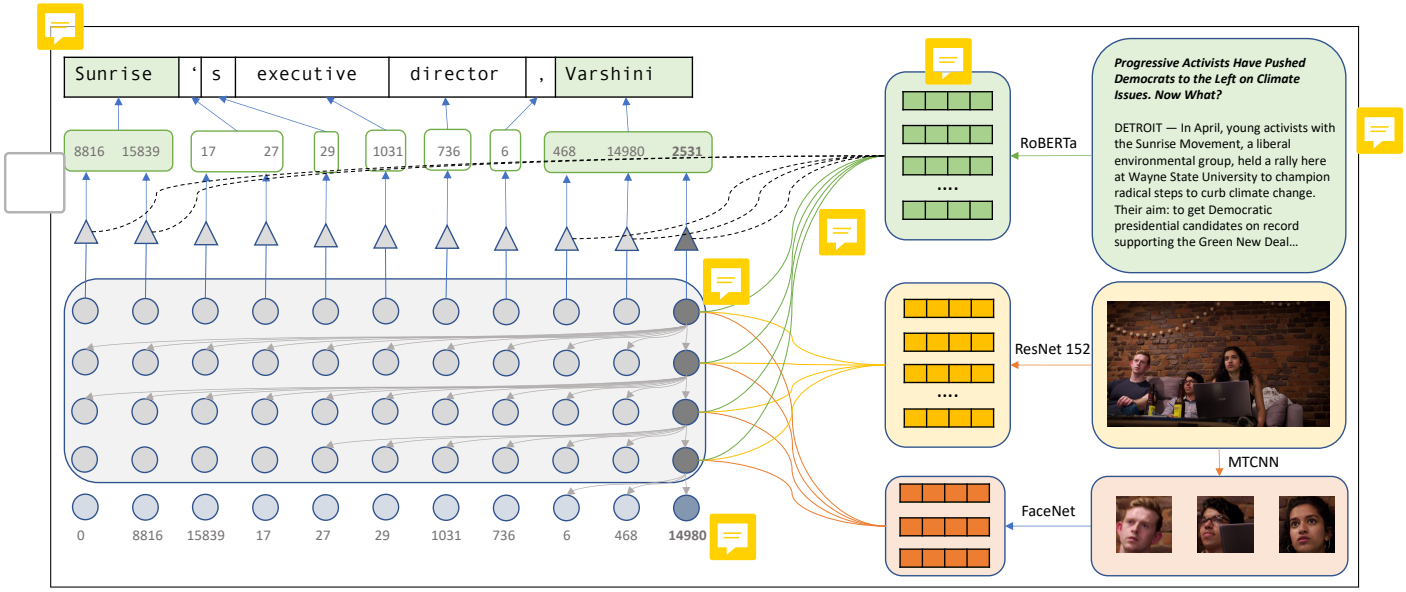

Figure 2: On the left is the decoder with four transformer blocks, which takes as input embeddings of byte-pair tokens (e.g. 14980 represents "arsh" in "Varshini"). The grey arrows inside the decoder indicate the convolutions in each block (we only show the convolutions in the final time step). After each convolution, we also attend to three domains on the right: article text (green lines), image patches (yellow lines) and faces (orange lines). The grey triangles are switches to determine if we should copy or not. If we copy, we attend to the article embeddings (dashed black lines) and copy the token with the highest attention score. The final decoder outputs are byte-pair tokens, which are then combined to form whole words and punctuations. Copied tokens and words are shaded in green.

entity labels, extracted using Google Cloud Vision API, as inputs to the model. In our work, we do not explicitly give the model a list of entities to appear in the caption, instead our model automatically identifies relevant entities from the provided news article.

BPE offers an elegant solution to handling an open vocabulary. To date the only image captioning work that uses BPE is [52], but they did not use it for rare named entities as these were removed from the captions during preprocessing. In contrast we explicitly examine the use of BPE for generating rare names and evaluate it in comparison to template-based methods.

In addition to attending to image patches, some captioning models also attend to object regions [46] and visual concepts [50, 22, 46], both of which are derived from the image itself. When attending to more than one modality, there are various strategies on how to combine embeddings such as addition, concatenation, and multivariate residual modules (MRMs) [18]. In our model we use the vector concatenation strategy and leave investigation of the more complex strategies, such as MMRs, to future work as they typically only yield minor performance improvements [46].

3. Model Architecture

Conceptually our model can be broken into two parts: encoding and decoding. The encoding part consists of a set of domain specific encoders for producing high level vector representations of images, faces and article text. The output of each encoder is a, potentially arbitrary, length set of fixed size vectors that represent the input. The decoder sequentially generates captions at the sub-word level by applying multi-headed attention over the sets of vectors from the encoders, and over a representation of the previously generated sub-word units. In practice there are a number details which allow us to train this large multi-faceted model on a single machine and achieve state-of-the-art performance. We have included some these details where appropriate and collected together those that do not fit elsewhere into Section 3.3.

3.1. Encoders

Our proposed model takes three types of inputs: image, faces, and article text. Each of these inputs is encoded into a set of vectors by domain specific encoders pre-trained on data from the matching domain.