News Image Caption Generation

Rahul Verma ai23mtech11008   Mohan KRK ai23resch14001   Manikanta Vallepu ai20btech11014

April 30, 2024

Abstract

This project explores the task of **news-image caption generation**, which involves producing descriptive sentences for news images based on associated articles. The dataset used in this project is extracted from the New York Times API. The approach, inspired by existing literature, integrates image and text features using commonsense knowledge reasoning and multimodal context reasoning. By combining information from the articles with objects and attributes extracted from the images, the project constructs a multimodal news context that is processed to generate news-image captions. This project demonstrates the potential to improve the understanding of news images through automated captioning.

1 Introduction

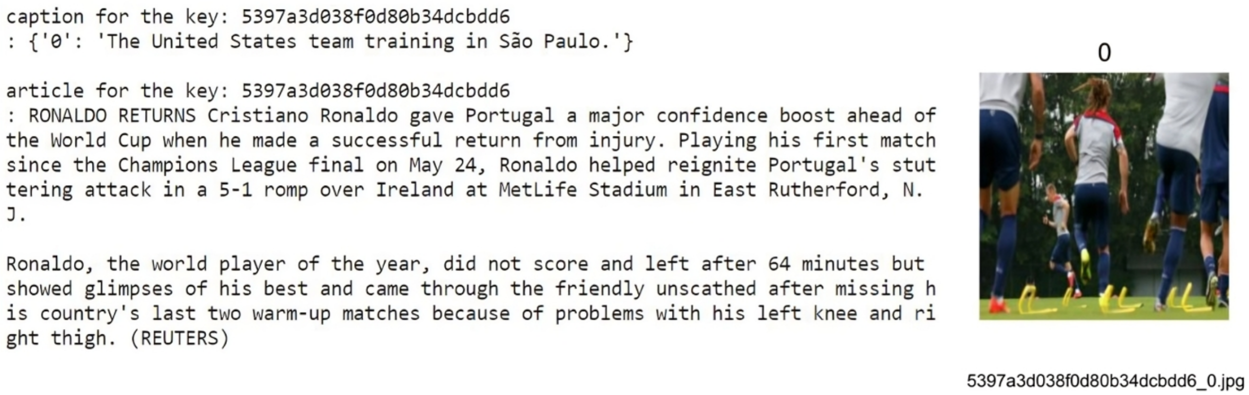
Image captioning, the automatic generation of natural language descriptions from images, has long been a focus for both the Computer Vision (CV) and Natural Language Processing (NLP) communities. This task offers practical applications such as automatic image indexing and poses a challenge in image understanding, including recognizing objects, their relationships, and scenes.

This project tackles the more advanced task of generating news image captions, where a descriptive sentence is created for a news image based on its associated article. News image captioning is more complex than traditional image captioning as it requires an accurate description of the named entities within the image and a thorough understanding of both the visual content and the context provided by the accompanying text.

In this project, we implement a method for news image caption generation based on existing literature that integrates both image and text features. Our approach demonstrates the potential to enhance news image comprehension through automated captioning. By combining data from the New York Times API with advanced models such as FLAN-T5 and CLIP, and using a GPT-2 decoder, our objective is to produce accurate and informative captions for news images.

2 Dataset for News-Image Captioning

For this study, we utilized the New York Times API to obtain a dataset comprising images, captions, and articles. We specifically curated a subset of the data, extracting 50,000 entries spanning the years 1999-2019. The dataset was stored in a JSON file format and organized by a unique identifier (key ID) to maintain consistency across the image and text data, which were stored separately.



```
caption for the key: 5397a3d038f0d80b34dcbdd6
: {'0': 'The United States team training in São Paulo.'}

article for the key: 5397a3d038f0d80b34dcbdd6
: RONALDO RETURNS Cristiano Ronaldo gave Portugal a major confidence boost ahead of
the World Cup when he made a successful return from injury. Playing his first match
since the Champions League final on May 24, Ronaldo helped reignite Portugal's stut
tering attack in a 5-1 romp over Ireland at MetLife Stadium in East Rutherford, N.
J.

Ronaldo, the world player of the year, did not score and left after 64 minutes but
showed glimpses of his best and came through the friendly unscathed after missing h
is country's last two warm-up matches because of problems with his left knee and ri
ght thigh. (REUTERS)
```

Figure 1: Overview of the dataset for a given key ID

3 Approach and Methodology

Given a news image I and its associated article A, our goal in news image captioning is to generate a descriptive sentence $S = \{s_1, s_2, \ldots, s_T\}$. Figure [2] presents the model architecture utilized for generating captions for the image I. The model comprises two key modules: commonsense knowledge reasoning and multimodal context reasoning. Let us understand the working of these modules in more detail.

3.1 Commonsense Knowledge Reasoning

For the captioning problem, we have the article A. Our goal is to extract as much information as possible from this article so that the generated caption is not generic and is aligned according to the article and image. For this purpose, we want to find all the *named entities* and the associated *entity context*. The *entity context* refers to the paragraph form in which the *named entity* is located. In the next step, we want to feed this knowledge prompt into a pre-trained model to get *commonsense knowledge reasoning*. Pre-trained model has shown significant improvement over the years and hence are capable for our task.

- **Knowledge prompt construction:** First step is to generate knowledge prompt. Knowledge problem has three parts: *instructions*, *named entity* and *entity context*. For example, our knowledge prompt will look like this, "Generate some knowledge about the named entity: {*named entity*} based on the news context: {*news context*}".

More formally, we define knowledge prompt p_n as,

$$p_n = [t; e_n; c_n]$$

where, t refers to the instruction, e_n refers to the named entity, and c_n is the entity context.

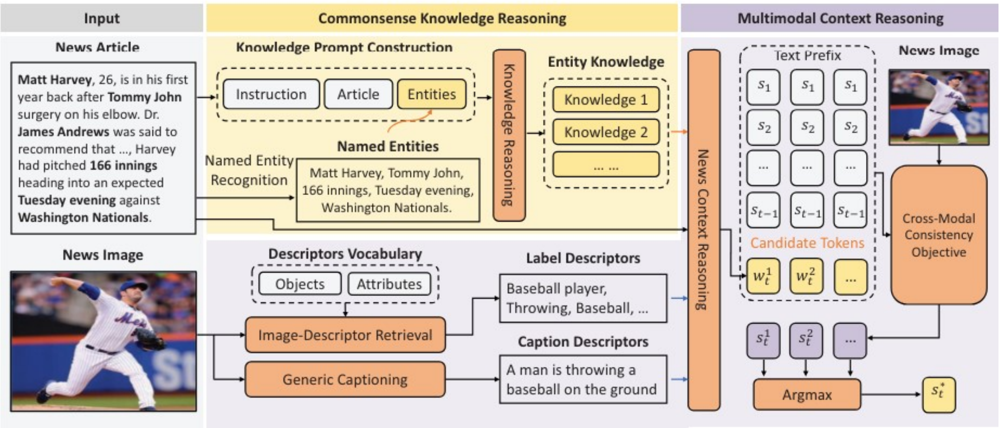

Input
Commonsense Knowledge Reasoning
Multimodal Context Reasoning
News Article
Matt Harvey, 26, is in his first year back after Tommy John surgery on his elbow. Dr. James Andrews was said to recommend that ..., Harvey had pitched 166 innings heading into an expected Tuesday evening against Washington Nationals.
Knowledge Prompt Construction
Instruction
Article
Entities
Named Entity Recognition
Named Entities
Matt Harvey, Tommy John, 166 innings, Tuesday evening, Washington Nationals.
Knowledge Reasoning
Entity Knowledge
Knowledge 1
Knowledge 2
... ...
News Context Reasoning
Text Prefix
s_1
s_2
...
s_{t-1}
Candidate Tokens
w_t^1
w_t^2
...
News Image
Cross-Modal Consistency Objective
s_t^1
s_t^2
...
Argmax
s_t^*
News Image
Descriptors Vocabulary
Objects
Attributes
Image-Descriptor Retrieval
Generic Captioning
Label Descriptors
Baseball player, Throwing, Baseball, ...
Caption Descriptors
A man is throwing a baseball on the ground

Figure 2: Model framework for news-image captioning

- **Get Knowledge Reasoning:** In the next step, we feed the prompt generated in the first step into a pre-trained model to generate knowledge reasoning. Knowledge assertions K can be represented as

  $$K = \{k_n : k_n \sim p(l_k|p_n), n = 1...N\}$$

  where k_n denotes the knowledge statement for the n-th context-aware knowledge prompt.

We have used **FLAN-T5** model to get commonsense knowledge reasoning which has been fine-tuned using instruction-tuning data from the FLAN (FlipDA Instruction-Aware Language model) corpus.

3.2 Multimodel Context Reasoning

Before generating caption for image I, we want to leverage the information contained in the image like labels, attributes, and different actions that are being performed in the image. For this, we decompose image into two steps, Word-level descriptors and sentence-level descriptors.

- **Word-level descriptor:** In this step, we retrieve the objects and attributes that best describe the image. Different objects and attributes has taken from *OpenImage* and *Visual Genome* attribute list. We first extract text embeddings and image embeddings using an advanced multimodel encoder, **CLIP**. Then we compute cosine similarity to get top-5 objects and attributes.

- **Sentence-level descriptor:** We want to get a generic caption of the image using a pre-trained captioning model. For that, we use **BLIP** model. It helps us to get an overview of the image within a sentence. Typically, we choose top-2 captions generated by this model.

We have gathered all essential features and contexts from both the image and article. Next, we combine the article (A), knowledge reasoning (K), word-level descriptors (L), and sentence-level descriptors ($\hat{S}$) to form a multimodal news context (C):

$$C = [A, K, L, S]$$

In the final step, we input the multimodal news context (C) into a language model to generate news-image captions. We used **GPT-2** as our language model for this purpose.

Algorithm 1 News-Image Caption Generation

```
Require: News image $I$ and article $A$
Ensure: Generated caption $S$
Commonsense Knowledge Reasoning
Extract named entities and context from $A$.
Construct knowledge prompt $p_n$.
Obtain commonsense reasoning $K$ from pre-trained model with $p_n$.
Multimodal Context Reasoning
Extract word-level descriptors from $I$ using multimodal encoder.
Extract sentence-level descriptors using captioning model.
Combine $L$ and $\hat{S}$.
Multimodal News Context Construction
Combine $A$, $K$, $L$, and $\hat{S}$ to form multimodal context $C$.
Caption Generation
Input $C$ into language model to generate $S$.
return $S$.
```

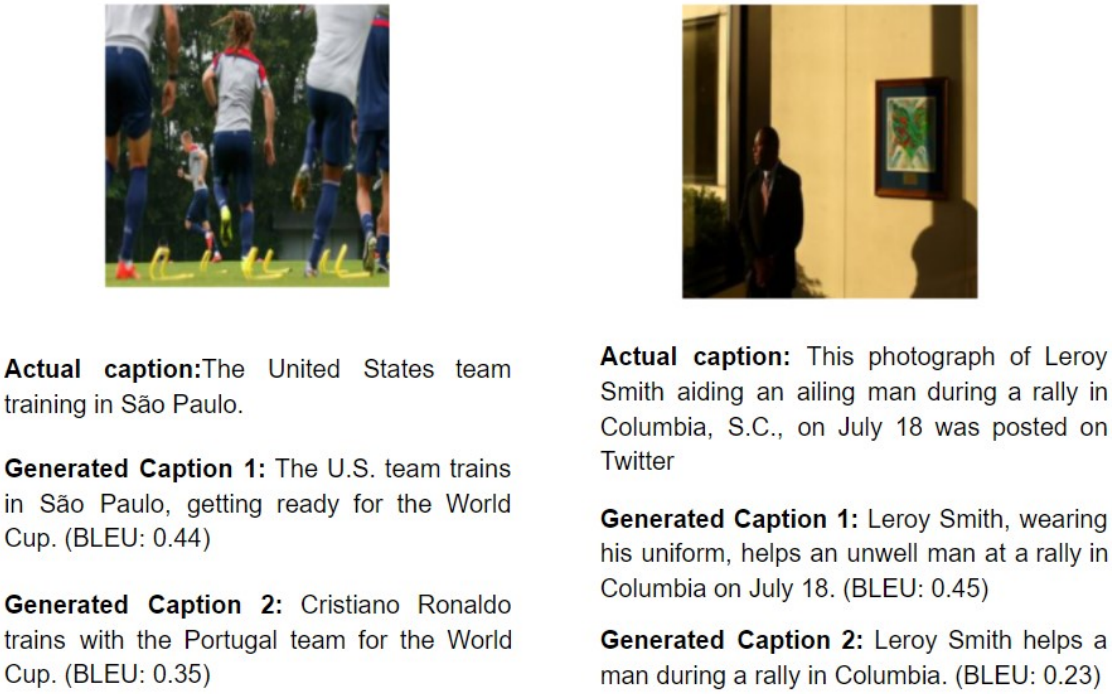
4 Results

Due to limited GPU capacity, we faced challenges in generating captions for images. To optimize the code's performance and functionality, we applied the following settings:

- The article text was truncated to the first 100 tokens for processing.
- We randomly selected 5 knowledge assertions from the FLAN-T5 model.
- Only the top 3 word-level descriptors were used.
- A beam size of 3 was set for the GPT-2 model.

With these settings, we generated two captions for each image. To evaluate the quality of the generated captions, we used the BLEU score as our metric. Figure 3 compares the actual caption with the generated captions for the image in the first column and includes the BLEU score for these captions.

To assess the overall quality of the model, we generated captions for 50 keys and calculated an aggregate BLEU score. The aggregate BLEU score we found was **0.38**.



Actual caption: The United States team training in São Paulo.

Generated Caption 1: The U.S. team trains in São Paulo, getting ready for the World Cup. (BLEU: 0.44)

Generated Caption 2: Cristiano Ronaldo trains with the Portugal team for the World Cup. (BLEU: 0.35)

Actual caption: This photograph of Leroy Smith aiding an ailing man during a rally in Columbia, S.C., on July 18 was posted on Twitter

Generated Caption 1: Leroy Smith, wearing his uniform, helps an unwell man at a rally in Columbia on July 18. (BLEU: 0.45)

Generated Caption 2: Leroy Smith helps a man during a rally in Columbia. (BLEU: 0.23)

Figure 3: Comparison of Actual and Generated Captions with BLEU Scores

5 Conclusion

In this project, we investigated news-image caption generation using a method that combines image and text features. We generated two captions per image and evaluated them with the BLEU score, achieving an aggregate score of 0.38.

While there is room for improvement, our approach offers a promising start toward enhancing the quality of automated news-image captioning. Future work can focus on refining the model and exploring additional strategies to improve performance.

References

[1] Y. Wang et al., ”Knowledge Prompt Makes Composed Pre-Trained Models Zero-Shot News Captioner,” 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 2023, pp. 2879-2884, doi: 10.1109/ICME55011.2023.00489.

[2] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797-812, April 2013, doi: 10.1109/TPAMI.2012.118.

[3] A. F. Biten, L. Gomez, M. Rusiñol and D. Karatzas, "Good News, Everyone! Context Driven Entity-Aware Captioning for News Images," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 12458-12467, doi: 10.1109/CVPR.2019.01275.