Deep Learning-Based Image Captioning for Food Datasets

Miren Samaniego
Barcelona, Spain
mirensamaniego.ikas@gmail.com

Laura Salort
Barcelona, Spain
laura.salort01@estudiant.upf.edu

Lucía Arribas Revilla
Barcelona, Spain
luciaarribasrevilla@gmail.com

Júlia Bujosa Moya
Barcelona, Spain
julbujmoy@gmail.com

Abstract

Image captioning is a fundamental task in computer vision and natural language processing, aiming to generate meaningful textual descriptions from images. This project explores deep learning-based approaches to image captioning, specifically applied to food datasets. We utilize a hybrid encoder-decoder architecture, where a ResNet-based CNN extracts visual features, and a GRU-based RNN generates textual descriptions. The dataset consists of food images paired with their respective dish titles, preprocessed and divided into training, validation, and test sets (80-10-10).

To improve performance, we experiment with different encoders (ResNet-18, ResNet-50, VGG-16) and various text representation levels (word-level and wordpiece-level). The model's performance is evaluated using standard NLP metrics, including BLEU, METEOR, and ROUGE. Our results demonstrate the effectiveness of deep learning models in generating accurate image captions for food-related datasets, highlighting key challenges such as handling out-of-vocabulary words and improving contextual consistency.

1. Introduction

Image captioning combines the fields of computer vision and natural language processing to automatically generate descriptive sentences from visual inputs. It has applications in accessibility, content generation, and automated image understanding. The rise of deep learning has enabled significant advancements in this field, with encoder-decoder architectures being the most widely used approach.

This project focuses on applying image captioning techniques to food datasets, aiming to generate accurate dish names based on image content. Unlike generic image captioning tasks, food image captioning presents unique challenges such as high intra-class similarity and subtle visual differences between dishes. To tackle this, we implement a deep learning-based model leveraging a CNN for feature extraction and an RNN for text generation.

The study follows a structured pipeline, including dataset preprocessing, model training, hyperparameter tuning, and evaluation using standard NLP metrics. The contributions of this work include experimenting with different encoder architectures (ResNet-18, ResNet-50, VGG-16), different decoder architectures (LSTM, xLSTM), exploring text representation levels (word-level vs. wordpiece-level), and analyzing the impact of various training strategies. Our findings provide insights into the challenges and potential improvements for food image captioning systems.

2. Related Work

Image captioning has evolved significantly in the last decade, driven by advances in deep learning and multimodal understanding. Early approaches relied on rule-based and retrieval methods, but the introduction of deep neural networks has led to significant advances. This section provides an overview of the main advances in image captioning, focusing on deep learning techniques, encoder-decoder architectures and evaluation methodologies.

2.1. Early Approaches to Image Captioning

Initial image captioning methods were mainly based on retrieval strategies, in which textual descriptions were matched to images using similarity measures. Template-based methods also played an important role, as they were based on predefined sentence structures with detected objects and attributes [2]. Although these approaches produced syntactically correct sentences, they lacked the flexibility to generate diverse and contextually relevant captions.

2.2. Deep Learning-Based Image Captioning

The introduction of deep learning revolutionized image captioning by enabling end-to-end learning of visual and linguistic representations. The standard framework consists of

an encoder-decoder architecture, in which a convolutional neural network (CNN) extracts the visual features and a recurrent neural network (RNN), such as long-term memory (LSTM) or controlled recurrent units (GRU), generates the text [3]. The seminal "Show and Tell" model by Vinyals et al. demonstrated the effectiveness of this approach, using a CNN to encode images and an LSTM to produce captions [2].

Recent advances have introduced attentional mechanisms to improve the alignment between visual features and generated words. Soft and hard attentional models selectively focus on different regions of the image, which improves the accuracy of captions [3]. More recently, Transformer-based models have outperformed RNNs by leveraging self-attenuation mechanisms to model long-range dependencies between words and visual elements [1].

2.3. Variants and Specialized Approaches

Several adaptations to specific domains have been studied, such as captioning of medical images and captioning of food images. In these cases, models must handle specialized vocabularies and subtle visual differences. For example, in food image captioning, distinguishing between visually similar dishes is a major challenge [1]. Some approaches incorporate external knowledge, such as ingredient lists or nutritional information, to improve the quality of the captions

Another emerging trend is the integration of reinforcement learning to optimize caption generation based on evaluation metrics such as BLEU and METEOR. Reinforcement learning allows fine-tuning models by rewarding captions more closely resembling human captions [3].

2.4. Evaluation Metrics and Datasets

The evaluation of image captioning models requires standardized metrics to compare the generated captions with human references. Among the most commonly used metrics are BLEU, METEOR, ROUGE and CIDEr, each of which captures different aspects of linguistic quality [3]. However, these automated metrics do not always perfectly match human opinion, leading to further research to improve evaluation methodologies.

Datasets play a crucial role in model development, with MSCOCO and Flickr30k being the most commonly used benchmarks [1]. For food image captioning, specialized datasets containing annotated dish names and ingredient descriptions are needed to train effective models.

2.5. Challenges and Future Directions

Despite the remarkable progress made, image captioning still presents several challenges. Managing out-of-vocabulary words, improving contextual coherence, and generating more diverse and natural captions are pending research problems [2]. In the future, it will be necessary to incorporate multimodal knowledge graphs, take advantage of large pre-trained vision-language models, and refine evaluation techniques to better reflect human preferences.

3. Methodology

4. Experimental Settings

4.1. Datasets

4.2. Metrics

5. Results

6. Conclusions

References

[1] Lakshita Agarwal and Bindu Verma. From methods to datasets: A survey on image-caption generators. *Multimedia Tools and Applications*, 83:28077–28123, 2024. 2

[2] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *arXiv preprint arXiv:2107.06912*, 2021. 1, 2

[3] Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. Deep image captioning: A review of methods, trends, and future challenges. *Neurocomputing*, 546:126287, 2023. 2