Image Captioning for Food Datasets: A Deep Learning Approach

Agustina Ghelfi
Universitat Autònoma de Barcelona
Barcelona, España
agustina.ghelfi@autonoma.cat

María José Millán
Universitat Autònoma de Barcelona
Barcelona, España
mariajose.millan@autonoma.cat

Laila Aborizka
Universitat Autònoma de Barcelona
Barcelona, España
lailamohamed.aborizka@autonoma.cat

Abstract

1. Introduction

Image captioning is a fundamental task in computer vision and natural language processing that involves generating descriptive text based on visual input. It combines deep learning techniques from both domains, requiring models to extract meaningful visual features from images and translate them into coherent and contextually relevant text. Over the years, various approaches have been developed, ranging from traditional encoder-decoder architectures to more advanced methods incorporating attention mechanisms and transformer-based models.

The objective of this project is to explore different methodologies for image captioning, starting from the basics and gradually incorporating more sophisticated techniques. The project follows a structured learning path, beginning with data preprocessing and baseline models, and advancing towards modifications and improvements that align with state-of-the-art approaches. This structured approach allows us to assess model performance both quantitatively and qualitatively while iteratively improving the results.

For this study, we use the Food Ingredients and Recipes Dataset, which consists of 13,582 images of various dishes. The primary goal is to develop models capable of predicting dish titles based on image inputs. The dataset is divided into 80% training, 10% validation, and 10% test, ensuring a robust evaluation framework.

Throughout this project, we aim to gain deeper insights into the key components that contribute to effective image captioning systems and push the boundaries of traditional methods by exploring more advanced techniques.

2. Related work

Image captioning is a multidisciplinary research area that intersects computer vision and natural language processing (NLP) to generate textual descriptions of images. Over the years, various approaches have emerged, evolving from handcrafted rules and retrieval-based methods to modern deep learning-based architectures.

Early methods primarily relied on template-based and retrieval-based techniques, where captions were either generated using predefined sentence structures or retrieved from a dataset based on image similarity. These approaches lacked flexibility and failed to generalize well to unseen images [2].

With the advent of deep learning, encoder-decoder architectures became the dominant paradigm. The first deep learning-based models utilized Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for sequential text generation [3]. While these models significantly improved caption quality, they struggled with long-range dependencies and lacked a mechanism to focus on specific image regions.

To address this, attention mechanisms were introduced, allowing models to dynamically focus on relevant parts of an image while generating each word. Notable variants include soft attention, which learns a probabilistic weight distribution over image regions, and hard attention, which stochastically selects regions at each time step. More recently, self-attention and transformer-based architectures, such as Vision Transformers (ViTs) and multimodal transformers, have further improved captioning performance by capturing long-range dependencies and integrating vision

and language more effectively [1, 2].

Another important development in image captioning is the use of reinforcement learning techniques, such as self-critical sequence training (SCST), which optimizes non-differentiable evaluation metrics like BLEU, CIDEr, and METEOR. These approaches enhance fluency and coherence beyond traditional maximum likelihood estimation [3]. Additionally, some works integrate graph-based representations and external knowledge bases to enrich captioning models with structured semantic relationships, leading to more contextually aware descriptions.

The field has also expanded into domain-specific applications, particularly in medical image captioning. Unlike general image captioning, where datasets such as MS COCO and Flickr30k are commonly used, medical image captioning requires specialized datasets like IU X-ray and MIMIC-CXR. These domain-specific models incorporate structured medical knowledge and leverage hierarchical learning techniques to generate accurate and clinically meaningful reports [1, 3].

Despite these advancements, several challenges remain. Current models still struggle with generating captions that are truly human-like in terms of coherence, diversity, and commonsense reasoning. Many image captioning models exhibit biases toward frequent objects and fail to describe rare or unseen entities effectively. Furthermore, evaluation metrics for image captioning remain a challenge, as automated scores often do not fully capture the quality of human-like descriptions [2].

Future research directions include enhancing multimodal pretraining techniques, leveraging large-scale vision-language models, and improving explainability in caption generation. The integration of external knowledge graphs, commonsense reasoning, and more sophisticated linguistic structures will also be crucial in advancing the field further.

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*, 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. 1, 2