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 applied to image captioning, specifically for food datasets. The first part of the project explores encoder-decoder architectures, where a ResNet-based CNN extracts visual features, and an RNN generates textual descriptions, evaluating different architecture combinations and text representations. The second part explores more advanced methods, with Visual Transformers to extract features from the images and Large Language Models to generate captions. Model performance is evaluated using standard natural language processing metrics, including BLEU, METEOR, and ROUGE. The obtained results demonstrate the potential of deep learning models, specially transformer-based, in generating accurate image captions, highlighting key challenges such as handling out-of-vocabulary words and improving contextual consistency.

1. Introduction

Image captioning combines computer vision and natural language processing to generate descriptive sentences from images. This project focuses on food image captioning, a task complicated by high intra-class similarity and subtle visual differences. We employ deep learning models, initially using CNNs for feature extraction and RNNs for text generation.

In the first phase, we explored different encoder architectures (ResNet-18, ResNet-50) and decoder models (LSTM, GRU), evaluating text representation levels and training strategies. In the second phase, we extended our approach by integrating Transformer-based architectures and large language models (LLMs). We experimented with the ViT-GPT2 model, comparing pretrained and fine-tuned variants. Additionally, we tested multimodal DeepSeek-VL and fine-tuned smaller Llama decoders (Llama 3.2-1B, Llama 3.2-3B) using LoRA with a frozen ViT encoder.

To systematically evaluate model performance, we conducted quantitative assessments using BLEU-1, BLEU-2, ROUGE-L, and METEOR scores. Furthermore, we analyzed the impact of fine-tuning strategies and encoder-decoder configurations on the quality of generated captions. The findings provide valuable insights into the effectiveness of Transformer-based models compared to traditional CNN-RNN approaches, highlighting potential improvements in food image captioning systems.

2. Related Work

Image captioning has evolved significantly in the last decade, driven by advances in deep learning and multimodal learning. Early approaches relied on rule-based and retrieval methods, but the advent of deep neural networks has enabled powerful end-to-end systems capable of generating diverse, context-aware captions. This section outlines key developments in the field, focusing on encoder-decoder architectures, attention mechanisms, and recent trends involving transformer-based and multimodal models.

2.1. Early Approaches to Image Captioning

Initial methods for image captioning were based on retrieval strategies, in which textual descriptions were matched to images using similarity measures. Template-based techniques also played a foundational role, using predefined sentence structures filled with detected objects or attributes [7]. While effective at producing grammatically correct sentences, these approaches lacked flexibility and adaptability to unseen images or concepts.

2.2. Deep Learning-Based Architectures

The transition to deep learning introduced encoder-decoder architectures for image captioning. A typical pipeline uses a convolutional neural network (CNN) to extract visual features and a recurrent neural network (RNN), such as an LSTM or GRU, to generate descriptive text [10]. The "Show and Tell" model by Vinyals et al. [8] established this paradigm.

The introduction of attention mechanisms marked a turning point in caption quality. The "Show, Attend and Tell" model [9] allowed the decoder to focus dynamically on different image regions while generating each word. Subsequently, Transformer-based models replaced RNNs, leveraging self-attention to model long-range dependencies. These models demonstrated superior performance, particularly in fluency and coherence [1].

2.3. Vision Transformers and Multimodal Pretrained Models

More recent work employs Vision Transformers (ViT), which model images as sequences of patches and excel at capturing global image context. When combined with powerful language decoders such as GPT-2 or LLaMA, ViT-based architectures have shown impressive results. These systems benefit from better alignment between visual and textual modalities due to the self-attention mechanism used in both encoder and decoder.

Multimodal pretrained models like DeepSeek-VL [5]. combine large-scale vision and language training to support tasks such as image captioning, visual question answering, and grounding. These models accept both images and textual prompts as input, and have been trained on billions of vision-language tokens, enabling zero-shot or few-shot capabilities without fine-tuning.

2.4. Fine-tuning Strategies and Parameter Efficiency

Fine-tuning large models has become more efficient with methods such as LoRA (Low-Rank Adaptation), which injects a small number of trainable parameters into frozen pretrained networks. LoRA enables rapid adaptation to domain-specific datasets while keeping the majority of the model unchanged. This approach has been successfully applied to LLaMA models for image captioning tasks, allowing the use of compact yet powerful captioning pipelines with reduced computational cost.

2.5. Domain-Specific and Food Image Captioning

Specialized domains, such as food image captioning, pose unique challenges due to high intra-class visual similarity. Datasets like Food-101 [3] and Recipe1M+ [6] support this subfield by providing rich annotations. Recent approaches have integrated auxiliary information such as ingredients or nutritional data to improve caption quality in this context.

2.6. Challenges and Future Directions

Despite considerable progress, several challenges remain: generating diverse, human-like captions; addressing bias and hallucination; and improving alignment between visual content and language. Future directions include leveraging multimodal foundation models, incorporating external knowledge sources, and advancing interpretability and human-aligned evaluation methods.

3. Methodology

This section describes the models evaluated in this project for image captioning. In the first part simpler models are explored, using convolutional neural networks (CNNs) to extract visual features from images, and recurrent neural networks (RNNs) generate text captions from these features in an encoder-decoder manner. In the second part the work is focused on transformer-based models.

3.1. Simple encoder-decoder models

Figure 1 illustrates the encoder-decoder architecture used for image captioning. The ResNet encoder extracts visual features, which are then processed by the GRU or LSTM decoder to generate a sequence of words forming the caption.

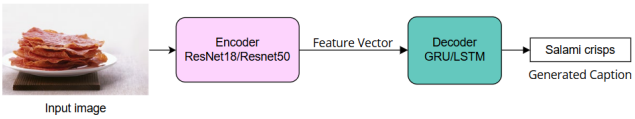

Figure 1. Encoder-Decoder architecture for image captioning.

3.1.1. Encoder: ResNet

For feature extraction, two different versions of ResNet were employed: ResNet-18 and ResNet-50. This architecture is known for its strong performance in capturing hierarchical image features. ResNet-18 is a lightweight version with 18 layers, providing a good balance between computational efficiency and performance. It is well-suited for tasks where computational resources are limited. ResNet-50 is a deeper model with 50 layers, capable of capturing more intricate and high-level features in images. It has more parameters than ResNet-18 and generally provides better performance in terms of feature extraction, albeit with higher computational costs.

Both ResNet-18 and ResNet-50 were pretrained on ImageNet and fine-tuned for the image captioning task. The final feature representations from the encoder are passed to the decoder for caption generation.

3.1.2. Decoder: GRU vs. LSTM

The text generation component utilizes two types of recurrent neural networks:

LSTMs (Long Short-Term Memory) are a type of RNN that are particularly well-suited for sequence generation tasks, as they are capable of modeling long-term dependencies in text. They use a more complex gating mechanism compared to GRUs, which makes them more powerful in certain tasks but at the cost of increased computational complexity.

GRUs (Gated Recurrent Unit) on the other hand are a more efficient alternative to LSTMs, with fewer parameters while still effectively capturing long-term dependencies in sequential data. GRUs are typically faster to train and can be more computationally efficient than LSTMs.

Both GRU and LSTM were used as decoders, and compared their performance with the ResNet-18 and ResNet-50 encoders to determine which configuration performed best for image captioning.

In both cases, the decoder receives the feature vector from the ResNet encoder and generates captions word by word, as seen in Figure 2. The hidden state of the encoder is used to initialize the hidden state of the decoder, and the sequence of words is generated using an embedding layer and recurrent layers. The output is a probability distribution over the vocabulary, with the most likely next word being selected at each time step.

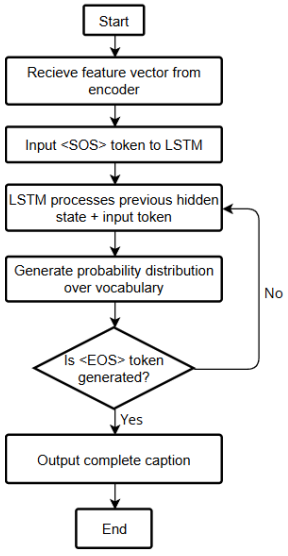

Figure 2. Illustration of the decoder process using an LSTM

3.2. Further Improvements

To enhance the baseline model performance several improvements were explored. First, different text representation levels were used, such as character, word-piece and word. Secondly, different number of layers for the decoders were used to study its effect on text generation performance. Another improvement was the implementation of attention mechanisms to allow the model to focus more on relevant image regions during caption generation. Furthermore, teacher forcing was introduced during training to stabilize the learning process by feeding the correct word from the previous time step into the model at each step, which could help the model learn faster and steadily while preventing error accumulation during early stages. Finally, beam search was used during inference to improve the quality of generated captions by exploring multiple caption hypotheses during generation, and selecting the most promising ones based on cumulative probability.

3.3. Transformer-based models

In the second phase of the project, state-of-the-art models for image captioning were evaluated, focusing on multimodal architectures that combine powerful visual and textual components. The following models were tested:

3.3.1. ViT-GPT2

This model combines a Vision Transformer (ViT-B/16) as the image encoder with GPT-2 Small as the language decoder [4], see Figure 3. The ViT encoder processes input images by dividing them into fixed-size patches and projecting them into a latent space using self-attention. This architecture allows it to capture long-range dependencies and contextual features, which are critical for effective visual understanding.

On the decoding side, GPT-2 is a Transformer-based language model that generates text autoregressively. It uses masked self-attention to ensure that each token attends only to previous ones, enabling coherent and contextually rich text generation. In this task, GPT-2 takes the high-level features extracted by ViT and produces natural language captions that describe food images. The ViT-GPT2 architecture combines a Vision Transformer for visual understanding with a pretrained language model for text generation.

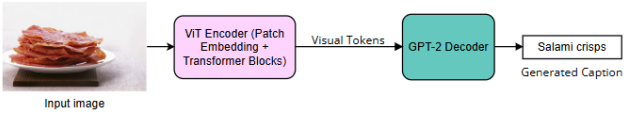

Figure 3. ViT encoder with GPT-2 decoder for image captioning. Visual tokens are projected to the GPT-2 embedding space.

Multiple training strategies were explored:

- Freezing the ViT and training only the GPT-2 decoder.
- Freezing the GPT-2 and training only the ViT encoder.
- Fine-tuning both components jointly.

Additionally, some techniques were applied during finetuning to improve the performance, apart form those mentioned

in 3.2 of teacher forcing and beam search. A repetition penalty was introduced, to aid in generating more diverse and coherent captions, as well as a no-repeat n-gram size constraint to avoid repeated phrases. Scheduled sampling was introduced during training to mitigate exposure biased of using teacher forcing, by gradually transitioning from using the ground truth tokens to its own predictions as inputs. Another approach that explored was combining the cross-entropy loss with a reinforcement learning (RL) loss derived from self-Critical Sequence training (SCST), computed as the difference in rewards between a baseline caption and a sampled caption. The latter are sampled during the RL updates using nucleus sampling (top-p), allowing the generation of diverse candidate sequences compared to the ones obtained with beam search. This is expected to allow the model to produce higher-reward outputs during inference. To compute the rewards, the METEOR metric was used.

3.3.2. ViT with LLaMA 3.2 (1B and 3B)

: Given that LLaMA is not multimodal, the model utilizes a frozen ViT encoder to extract visual features from resized input images (224×224), which are then passed through a projection layer to align them with the LLaMA embedding space. The language decoder consists of LLaMA 3.2 models with 1B and 3B parameters, fine-tuned using LoRA (Low-Rank Adaptation) to enable lightweight adaptation while preserving the core pretrained knowledge. Image captions and textual prompts (e.g., *"The name of the dish in the image is:"*) are tokenized and fed to the decoder alongside visual tokens, enabling conditional generation of descriptive captions. The overall pipeline was adapted from the implementation available at [2], with modifications for our specific dataset and training setup. Figure 4 shows the pipeline used to fine-tune LLaMA models with LoRA while keeping the ViT encoder frozen.

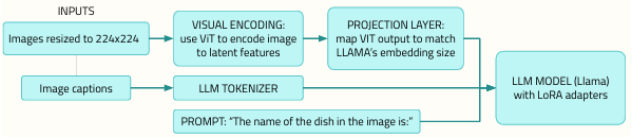

Figure 4. Architecture combining a frozen ViT encoder and LLaMA 3.2 (1B or 3B) decoder fine-tuned with LoRA adapters.

3.3.3. DeepSeek-VL (7B Chat)

: A powerful large multimodal model designed for visual-language understanding and generation. It uses a hybrid vision encoder composed of SigLIP-L and SAM-B, allowing it to process both appearance and spatial information from images. The model is built upon DeepSeek-LLM-7B-Base, a language model trained on approximately 2 trillion text tokens, and has been further trained on 400 billion vision-language tokens to enhance its multimodal capabilities. DeepSeek-VL accepts both images and textual prompts as input and was used in inference mode only, without additional fine-tuning.

A detailed comparison and analysis of the performance of these models is provided in Section 5.

4. Experimental Settings

4.1. Datasets

We used the Food Ingredients and Recipes Dataset available on Kaggle. The dataset comprises various food images paired with their respective recipe titles and descriptions. To facilitate model training, the dataset was divided into three subsets: 80% for training, 10% for validation, and 10% for testing.

4.1.1. Data Cleaning and Preprocessing

The dataset included not only relevant food images but also a significant number of irrelevant samples. To ensure a high-quality dataset, we performed several filtering steps:

- Removed images that contained large amounts of text, such as recipe book covers, promotional images with text overlays, and images featuring only textual descriptions of dishes (e.g., images with the word *PASTA* written in large letters).
- Excluded non-food-related images, such as those featuring only cutlery, windows, or other unrelated objects.
- Eliminated partial food images, where the food was only partially visible (e.g., showing just the plate without the food).
- Kept images featuring people eating, as these were considered relevant for the model's learning process.

Additionally, we observed that some images were duplicated but had different captions. This introduces noise into the dataset, as inconsistent captions for the same image can confuse the model. The following table shows the dataset sizes before and after cleaning:

State	Train size	Validation size	Test size
Before cleaning	10776	1348	1350
After cleaning	10309	1283	1297

Table 1. Dataset sizes before and after cleaning.

We considered two potential strategies to address the duplicated images issue: clustering similar images and assigning them multiple captions; and retaining only one caption per image, assuming the rest to be noise. These strategies were not implemented in this project but are noted as potential future improvements.

To automatically filter out text-heavy images, we applied a text detection model **pytesseract**, used for OCR tasks, with a specified threshold t. Images containing more than t text were removed from the dataset. Manual verification

was conducted to ensure completeness and accuracy of the cleaning process.

4.2. Metrics

We evaluated the models using standard Natural Language Processing metrics commonly used for image captioning:

BLEU-1, BLEU-2: Measures precision of n-grams. BLEU score assesses the overlap between n-grams of the generated caption and reference captions. The score ranges from 0 to 1, where 1 indicates a perfect match. It uses a modified precision metric, considering n-grams up to length 4 (BLEU-1 for unigrams, BLEU-2 for bigrams, etc.) and applies a brevity penalty to penalize short predictions.

$$\text{BLEU} = \min\left(1, \frac{\text{output\_length}}{\text{reference\_length}}\right) \cdot \left(\prod_{i=1}^{4} P_i\right)^{\frac{1}{4}}$$

ROUGE-L: Evaluates the longest common subsequence (LCS) between generated and reference captions. Unlike BLEU, which focuses on precision, ROUGE-L is based on recall. It captures how well the generated caption covers important sequences of words in the reference. A high ROUGE-L score indicates that the generated caption captures a significant portion of the reference caption's structure.

$$\text{ROUGE-L}_{F1} = \frac{2 \cdot P \cdot R}{P + R},$$

where P = precision and R = recall.

METEOR: Considers synonymy, stemming, and alignment between generated and reference captions. It computes a harmonic mean of precision and recall, giving higher weight to recall. Additionally, it uses WordNet synonym matching, stemming, and penalizes fragmented matches. This makes METEOR more robust to variations in wording compared to BLEU and ROUGE.

$$\text{METEOR} = (1 - \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}) \cdot \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R},$$

where m = nº of matched unigrams, ch = nº of chunks, $\alpha = 0.9$ (weighting factor), $\beta = 3$ (penalty shape), $\gamma = 0.5$ (penalty weight).

These metrics provide complementary perspectives on model performance. BLEU focuses on precise n-gram matches, ROUGE-L measures structural similarity, and METEOR addresses semantic similarity.

4.3. Implementation details

Regarding the implementation details, for the simple encoder-decoder architectures mentioned in 3.1, the Adam optimizer was used, with a learning rate of 0.001, the scheduler was ReduceLROnPlateau, with 5 warmup epochs, a batch size of 32, and a patience of 10 for early stopping. For the ViT-GPT2 the optimizer was AdamW with weight decay of 0.001 and a learning rate of $1e-5$, the scheduler was Cosine Annealing with $T\_max = 100$, $eta\_min = 1e-8$ and was then changed to CosineAnnealingWarmRestarts, with $T_0 = 10$, $eta_min = 1e-8$, when Reinforcement Learning was introduced to avoiding sharp drops in LR and stabilize training when combining CrossEntropy and Reinforcement Learning objectives. Additionally, patience was increased to 15. Finally, for the finetuning of both Llama 3.2 models the same optmizer and scheduler as with ViT-GPT2 without RL were used, reducing the patience for early stopping to 5. For the 1B model, a batch size of 6 was used, and the LoRA attention dimension (rank) was set to 8, with alpha parameter for scaling of 32. For the 3B model the batch size had to be reduced to 1 due to computational limitations, the LoRA rank was changed to 4, and also 2 to observe the impact in finetuning, both with an alpha of 16.

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] AnyModal. https://huggingface.co/AnyModal. 4

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. 2

[4] NLP Connect. vit-gpt2-image-captioning (revision 0e334c7), 2022. 3

[5] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 2

[6] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: a dataset for learning cross-modal embeddings for cooking recipes and food images. *arXiv preprint arXiv:1810.06553*, 2018. 2

[7] 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

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057, Lille, France, 2015. PMLR. 2

[10] 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