Automated Image Captioning Using Deep Learning

Deepika Hemant Tendulkar
George Mason University
dtendulk@gmu.edu

Sanjana Vegesna
George Mason University
svegesn@gmu.edu

Abstract

This project explores deep learning models for automated image captioning, a crucial task in bridging the gap between computer vision and natural language processing. We investigate the BLIP and BLIP-2 models, leveraging the COCO 2017 dataset, a rich resource for image captioning research. Our key results demonstrate the effectiveness of fine-tuning the BLIP model for improved performance and highlight the strong zero-shot capabilities of the BLIP-2 model, particularly when combined with the OPT-2.7B large language model, which offers a compelling balance between performance and inference speed. We achieve competitive scores on standard evaluation metrics, demonstrating the potential of these models for advancing the field of image captioning.

1. Introduction

1.1. Problem and Impact

Automated image captioning is a fundamental challenge in artificial intelligence, requiring a system to understand the visual content of an image and express that understanding in natural language. This task lies at the intersection of computer vision and natural language processing, demanding sophisticated techniques from both domains. The ability to automatically generate descriptive captions for images has far-reaching implications across various fields.

One of the most significant areas of impact is accessibility for visually impaired individuals. Image captioning technology can be integrated into assistive technologies, such as screen readers, to provide audio descriptions of images on web pages, social media, and other digital platforms, enabling users to access and understand visual information.

In the realm of information retrieval, image captioning can enhance the accuracy and efficiency of image search engines. By associating textual descriptions with images, it becomes possible to search for images based on their content, rather than relying solely on keywords or tags. This can be invaluable in applications such as content management, digital libraries, and e-commerce.

Furthermore, image captioning plays a crucial role in advancing human-computer interaction. It can facilitate more natural and intuitive communication between humans and machines, allowing systems to understand and respond to visual input in a way that is more aligned with human cognition. For example, in robotics, image captioning can enable robots to describe their surroundings, providing valuable information for navigation and task execution.

The development of robust and accurate image captioning systems represents a significant step towards achieving more general artificial intelligence, where machines can perceive, understand, and interact with the world in a manner that is more similar to humans.

1.2. Summary of approach

To tackle the image captioning task, we compared four vision-language models using the COCO dataset:

1. BLIP (Baseline): Zero-shot caption generation using the pretrained BLIP model without fine-tuning.

2. BLIP (Fine-tuned on COCO): The same BLIP model further trained on the COCO dataset to specialize its captioning capabilities.

3. BLIP-2 Flan-T5-XL: A multimodal encoder-decoder model integrating visual features with Flan-T5-XL as the text decoder.

4. BLIP-2 OPT 2.7B: A larger version using the OPT-2.7B language model for high-quality captioning in zero-shot mode.

All models were tested on a held-out set of COCO images using standard evaluation metrics such as BLEU, METEOR, ROUGE L, CIDEr, SPICE and ClipScore. The models were evaluated for both descriptive accuracy and semantic alignment with human-written captions.

1.3. Key results

Our experiments demonstrate that both fine-tuning and model architecture significantly influence image captioning performance. Fine-tuned models like BLIP consistently outperform their zero-shot counterparts, highlighting

the value of task-specific adaptation. Among the evaluated models, BLIP-2 OPT 2.7B delivered the strongest overall results, indicating that larger language models better capture nuanced visual-textual relationships. These results validate the effectiveness of combining powerful vision encoders with advanced language decoders in improving caption generation quality.

2. Approach

2.1. Background

Deep learning has revolutionized the field of image captioning, leading to significant advancements in the accuracy and fluency of generated captions.

Early approaches to image captioning often relied on Convolutional Neural Networks (CNNs) for visual feature extraction. CNNs, such as VGG and ResNet, were used to encode images into fixed-length vectors, representing the salient visual information. These feature vectors were then fed into Recurrent Neural Networks (RNNs), such as LSTMs or GRUs, which were responsible for generating the caption, word by word.

More recently, Transformer-based models have emerged as a powerful alternative to RNNs for sequence generation tasks, including image captioning. Transformers, with their attention mechanism, allow the model to weigh the importance of different parts of the input image when generating each word in the caption. This has led to improved caption quality and the ability to capture long-range dependencies in the text.

The BLIP model builds upon these advancements, utilizing a Vision Transformer (ViT) to encode images. ViT divides an image into a sequence of patches and processes them similarly to words in a sentence, allowing the Transformer architecture to be applied to visual data. The output of the ViT encoder is then fed into a BERT-style transformer, which is responsible for generating the caption. BLIP is pre-trained on a large dataset of image-text pairs, enabling it to learn strong associations between visual and linguistic information. This pre-training process allows the model to learn a general understanding of how images and language relate, which can then be fine-tuned for specific captioning tasks.

BLIP-2 takes a different approach, recognizing the increasing power and availability of large language models (LLMs). LLMs, pre-trained on massive amounts of text data, have demonstrated remarkable capabilities in generating coherent and fluent text, even with limited task-specific training. BLIP-2 leverages these powerful LLMs by introducing a lightweight Querying Transformer. This module learns to extract relevant visual features from a frozen image encoder and project them into the input space of the LLM. By freezing the image encoder and the LLM, BLIP-2 significantly reduces the computational cost of training, while still achieving strong performance.

2.2. Data sets

The dataset used for this project is the COCO (Common Objects in Context) 2017 dataset, which is widely used in image captioning tasks. It contains over 118,000 training images and 5,000 validation images, each annotated with five human-generated captions. These captions provide rich descriptions that capture the context, objects, and relationships within each image. For this project, subsets of the COCO validation set were used during both fine-tuning and evaluation phases to ensure consistency and manage computational resources effectively. The diversity and complexity of COCO make it ideal for benchmarking captioning models like BLIP and BLIP-2.

2.3. Implementation Details

The implementation was carried out using Python with PyTorch and the HuggingFace transformers library. The base model used for fine-tuning was Salesforce/BLIP, a vision-language encoder-decoder model pre-trained on large-scale datasets. The training and evaluation were performed in a Google Colab Pro+ environment with NVIDIA T4 GPU support.

We defined a custom PyTorch Dataset class to load image-caption pairs from the COCO dataset. We used 10,000 training images and 2000 validation images for the dataset. Each image was preprocessed using the BlipProcessor from HuggingFace, which handles image resizing, normalization, and tokenization of the captions. The training leveraged mixed precision (fp16=True) to reduce memory usage and increase throughput.

For fine-tuning BLIP, the Seq2SeqTrainingArguments and Seq2SeqTrainer classes from HuggingFace were utilized. Key configurations included a learning rate of 3e-5, batch size of 16, 3 training epochs, evaluation every 500 steps, and checkpointing every 1000 steps. The predict with generate flag ensured that caption generation was used during evaluation, not just label prediction.

For the BLIP-2 models (FLAN-T5-XL and OPT 2.7B), only zero-shot inference was performed due to their large size and memory requirements. Captions were generated by passing test images through the Blip2Processor and Blip2ForConditionalGeneration, without any further training. The COCO evaluation toolkit and pycocoevalcap were used to compute standard metrics like BLEU, METEOR, ROUGE L, CIDEr, and SPICE.

All models were evaluated on a common subset of the COCO validation dataset to ensure consistency in comparison.

2.4. Technical Approach and Innovation

In this project, we adopted a multi-faceted approach to explore and evaluate deep learning models for image captioning. Our approach is distinguished by its comprehensive model evaluation and enhancement strategy, which can be broken down into the following key elements:

1. **Model Selection and Setup:** We began by selecting a set of state-of-the-art models known for their strong performance in image captioning. This included BLIP (Bootstrapping Language-Image Pre-training) and BLIP-2. For BLIP-2, we chose two different large language models (LLMs) to evaluate their impact on performance: OPT 2.7B and Flan-T5-XL. We utilized the Hugging Face Transformers library and Salesforce's LAVIS library to streamline the process of model loading, inference, and caption generation. This allowed us to efficiently manage and experiment with these complex models.

   Model Architecture Summary:

   - **BLIP (Zero-Shot):**
     BLIP leverages a ViT-based (Vision Transformer) vision encoder to extract visual features, which are then fed into a BERT-style transformer for caption generation. It is pre-trained on 129 million image-text pairs, enabling it to generate captions in a zero-shot setting.
   - **BLIP (Fine-Tuned):**
     The BLIP model is fine-tuned on the COCO dataset to adapt it to the specific characteristics of this dataset. This fine-tuning enhances its performance, making it more suited for generating accurate captions for COCO images.
   - **BLIP-2 Flan-T5-XL:**
     BLIP-2 employs a Querying Transformer to interface between a frozen image encoder and the Flan-T5-XL large language model. This setup allows the model to generate captions based on the visual input, with Flan-T5-XL providing powerful language understanding for better caption generation.
   - **BLIP-2 OPT 2.7B:**
     Similar to BLIP-2 with Flan-T5-XL, BLIP-2 with OPT-2.7B also uses a Querying Transformer to interface between a frozen image encoder and the OPT-2.7B large language model for caption generation. However, OPT-2.7B offers faster inference times compared to Flan-T5-XL, making it a suitable choice for faster processing while still maintaining strong performance.

2. **Dataset and Preprocessing:** We employed the COCO 2017 dataset, a widely recognized benchmark for image captioning, for both training and evaluation. To effectively handle this dataset, we implemented a custom PyTorch Dataset class. This custom class was designed to parse the COCO 2017 annotations, extracting essential information such as image IDs, file names, and the multiple reference captions associated with each image. This preprocessing step was crucial for organizing the data in a format suitable for training and evaluating our models.

3. **Fine-Tuning:** To enhance the BLIP model's ability to understand image context and generate more descriptive and relevant captions, we performed fine-tuning on the COCO train2017 set. This process involved adapting the pre-trained BLIP model to the specific characteristics of the COCO dataset. To ensure training stability and optimize performance, we applied several techniques, including learning rate scheduling, dropout regularization, and caption-token alignment. These techniques helped to prevent overfitting, improve generalization, and ensure that the model learned to associate the correct words with the corresponding visual elements in the images.

   Training Details (BLIP Fine-Tuning):

   - **Learning rate scheduling** to stabilize convergence, with values ranging from 3e-5 to 5e-4.
   - **Dropout regularization** to prevent overfitting and enhance model generalization.
   - **Caption-token alignment** to improve the association between visual features and corresponding words.
   - **Mixed precision (FP16)** was enabled to achieve 2x faster training and reduce memory usage.
   - **Epochs:** 3–5
   - **Warmup Steps:** 500

4. **Innovation in Evaluation:** We expanded beyond traditional evaluation metrics in image captioning, such as BLEU and METEOR, by incorporating more advanced metrics that offer a deeper understanding of model performance. Notably, we introduced **CLIP-Score**, which evaluates the alignment between the vision and language modalities by computing the cosine similarity between the image and text embeddings generated by the CLIP model. We implemented a custom `calculate_clip_score()` function using OpenAI's CLIP model to compute this metric. Additionally, we used **SPICE**, which focuses on evaluating the semantic content of generated captions, providing a more detailed assessment of their quality.

Our evaluation methodology combined both conventional NLP metrics and innovative techniques to provide a comprehensive view of model performance. Traditional n-gram-based metrics like BLEU 1–4 and ROUGE-L measured surface-level similarity between generated captions and reference captions, while METEOR considered synonyms and stemming to improve flexibility. We also used **CIDEr** scoring to assess how well the generated captions align with human consensus.

Key Innovations:

- We conducted a controlled comparison between fine-tuned and zero-shot models, demonstrating that fine-tuning significantly improves all metrics. The BLIP Fine-Tuned model outperformed all other models in BLEU, METEOR, ROUGE, and SPICE.
- We also showed the architectural benefits of BLIP-2 over the BLIP baseline, especially in zero-shot mode, highlighting the improvements in performance.
- To ensure a fair comparison, we used the same COCO validation split for all models, providing consistency and reliability in the evaluation.

5. **Comparative Analysis:** To facilitate a comprehensive comparison of the models' performance, we created a detailed result table. This table presents the caption outputs generated by each model for the same set of input images, allowing for a direct comparison of their qualitative differences.

   We also analyzed how different evaluation metrics, such as CIDEr and CLIPScore, correlate with visual trends in the captions, such as their richness and brevity. This analysis helped us to gain insights into the strengths and weaknesses of each model and to understand how different metrics capture different aspects of caption quality.

3. Results

3.1. Evaluation Metrics

To assess the quality of the generated image captions, we used a comprehensive set of standard evaluation metrics commonly employed in image captioning research. These metrics evaluate various linguistic aspects such as n-gram overlap, semantic similarity, and syntactic fluency between the generated captions and the human-annotated ground truth captions. The metrics used are:

- **BLEU-1 to BLEU-4:**
  These metrics capture n-gram overlaps between generated and reference captions. BLEU-n measures the precision of n-grams (sequences of n words) in the generated caption compared to reference captions, with higher n indicating longer sequences of words matched.

- **METEOR:**
  This metric accounts for synonyms, stemming, and word order. METEOR addresses some limitations of BLEU by using more flexible matching criteria to improve alignment between generated and reference captions.

- **ROUGE-L:**
  ROUGE-L measures the longest matching sequence between the generated and reference captions, focusing on the longest common subsequence. It provides a balance between precision and recall in evaluating caption quality.

- **CIDEr:**
  This metric measures how well the generated captions align with human consensus. CIDEr evaluates the similarity between the generated and reference captions, with a higher emphasis on less frequent, more informative n-grams in the English language.

- **SPICE:**
  SPICE evaluates the semantic content of generated captions by representing them as semantic graphs. It compares the relationships between objects and their attributes, providing a more detailed assessment of scene-level semantics.

- **CLIPScore:**
  CLIPScore measures the alignment between the image and the generated caption in a high-dimensional embedding space learned by the CLIP model. It provides an assessment of visual-text alignment, highlighting how well the image content is captured by the generated caption.

These metrics were computed using the `pycocoevalcap` toolkit and custom scripts. They provide a multi-dimensional view of caption quality, balancing surface-level accuracy with deeper semantic understanding.

Metric	BLIP (Baseline)	BLIP (Fine-Tuned)	BLIP-2 Flan-T5-XL	BLIP-2 OPT 2.7B
BLEU-1	0.647	0.769	0.745	0.748
BLEU-2	0.522	0.611	0.586	0.613
BLEU-3	0.407	0.470	0.443	0.483
BLEU-4	0.313	0.357	0.330	0.375
METEOR	0.241	0.287	0.272	0.275
ROUGE_L	0.535	0.575	0.560	0.580
CIDEr	1.011	1.227	1.118	1.251
SPICE	0.184	0.220	0.216	0.217
CLIPScore	29.1	31.2	30.5	32.0

Figure 1. Evaluation metrics comparison of different captioning models

Key Observations

We evaluated the performance of four models on the COCO validation dataset using the standard image captioning metrics mentioned above. The models tested were: *BLIP (Baseline Zero-shot)*, *BLIP (Fine-Tuned)*, *BLIP-2 Flan-T5-XL*, and *BLIP-2 OPT 2.7B*. The key results are summarized as follows:

- **BLIP (Fine-Tuned):** Shows consistent improvement over the baseline model across nearly all metrics, demonstrating the effectiveness of fine-tuning on the COCO dataset.
- **BLIP-2 OPT 2.7B:** Achieves the highest scores in CIDEr, BLEU-3/4, and CLIPScore, suggesting it captures richer and more relevant image-text relationships due to its larger model size.
- **BLIP-2 Flan-T5-XL:** Performs competitively but slightly lags behind BLIP-2 OPT in most metrics, particularly in CIDEr and BLEU-4.
- **BLIP (Baseline Zero-shot):** Performs the weakest, especially on METEOR and SPICE, confirming the benefits of task-specific training.

These results validate that both fine-tuning and larger transformer architectures contribute significantly to improved captioning performance.

3.2. Qualitative Results Analysis

In addition to the quantitative evaluation metrics, a qualitative analysis was conducted to assess the semantic richness, contextual grounding, and fluency of the generated captions. This involved comparing the captions generated by four different models on the same set of images.

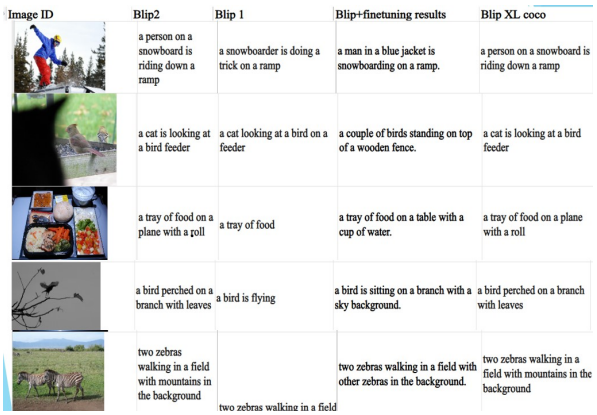

Figure 2. Qualitative Results Analysis

The table above compares the captions generated by BLIP, BLIP Fine-Tuned, BLIP-2 OPT 2.7B, and BLIP-2 Flan-T5-XL for a selection of images:

- **BLIP:** The generated captions were reasonable, capturing the core subject and action, but occasionally missing fine details. For example, in the image of the snowboarder, BLIP produced *"a snowboarder is doing a trick on a ramp,"* which is close but lacks specific details, such as the color of the jacket.
- **BLIP Fine-Tuned:** Showed noticeable improvement in caption quality, producing more descriptive and accurate captions. For the snowboarder, it enhanced the description to *"a man in a blue jacket is snowboarding on a ramp,"* providing richer context.
- **BLIP-2 OPT 2.7B:** Demonstrated solid performance but sometimes provided more general captions. For example, *"a bird is flying"* lacks specific details that could better describe the scene.
- **BLIP-2 Flan-T5-XL:** Produced the most detailed and accurate captions. For the snowboarder, it correctly described *"a person on a snowboard is riding down a ramp,"* showing a refined understanding of the image.

3.3. Conclusion

The combination of quantitative and qualitative results strongly supports the benefits of both fine-tuning and leveraging larger transformer architectures for improved image captioning. Fine-tuning (as seen with BLIP) significantly enhances the model's ability to generate more accurate and descriptive captions compared to a zero-shot approach. Furthermore, larger models like BLIP-2 OPT 2.7B demonstrate the capacity to capture more complex image-text relationships and align better with human consensus.

While BLIP-2 Flan-T5-XL shows promise in generating highly detailed captions, its slightly lower quantitative

scores compared to BLIP-2 OPT 2.7B warrant further investigation to understand the nuances of its performance across a larger dataset. These findings validate the importance of both fine-tuning and model architecture in advancing image captioning performance.

4. Related Work

Based on the article "Vision Transformers in Image Captioning Using Pretrained ViT Models" by Mobarak Inuwa, Vision Transformers (ViTs) have emerged as a highly effective architecture for image captioning tasks. The article highlights the use of pretrained ViT models for generating image captions by leveraging a VisionEncoderDecoderModel, which combines a vision transformer as the encoder with a GPT-2 model as the decoder. ViTs are advantageous due to their ability to process image data as sequences of patches, allowing the transformer architecture to operate on visual input similarly to how it works on textual data.

In this approach, the ViT model is fine-tuned for image captioning, and the GPT-2 model is used to generate descriptive captions from the encoded image features. The article also discusses the use of Hugging Face's pre-trained models for implementing image captioning, with applications extending beyond image captioning to areas like Optical Character Recognition (OCR), image detection, and deepfake identification.

The use of pretrained ViT models offers a practical solution for image captioning without the need to train models from scratch, providing an efficient and effective approach to leveraging vision transformers for multimodal tasks.

This approach aligns with recent advancements in image captioning using transformers, highlighting the growing interest in combining vision and language models to achieve accurate and contextually rich caption generation.

In the paper "Image and Text Features Extraction with BLIP and BLIP-2: How to Build a Multimodal Search Engine" by Enrico Randellini (2023), image captioning is explored using transformer-based models like BLIP and BLIP-2. These models combine Vision Transformers (ViTs) and large language models (LLMs) to efficiently generate captions. ViTs are used to encode images into embeddings, while a text transformer generates corresponding captions. BLIP-2 introduces a Querying Transformer (Q-Former), which bridges the image encoder and LLMs, reducing computational costs. Both models leverage multitask pre-training to align image-text pairs, enabling accurate caption generation. This architecture serves as a unified solution for multimodal tasks like image captioning and retrieval.

5. Resources

Software Libraries and Models Used

In this project, we leveraged several key software libraries and pre-trained models, primarily accessible through the Hugging Face ecosystem. The main resources we used included:

- **`transformers` Library:** Core library for accessing and utilizing the pre-trained BLIP models. It provided classes for model architectures, pre-trained weights, tokenizers, and processors, significantly simplifying our implementation.
  GitHub Repository: `https://github.com/huggingface/transformers`
- **`torch` (PyTorch):** Deep learning framework used to run and build the BLIP models.
  Website: `https://pytorch.org/`
- **`PIL` (Pillow):** Used for loading and manipulating image files.
  Website: `https://python-pillow.org/`
- **`numpy`:** Used for numerical computing, data array handling, and post-processing model outputs.
  Website: `https://numpy.org/`
- **`tqdm`:** Displayed progress bars during caption generation.
  GitHub Repository: `https://github.com/tqdm/tqdm`
- **`json`:** Python's built-in library used for parsing COCO annotations in JSON format.
- **`os`:** Python's built-in library used for constructing file paths and interacting with the file system.
- **`pycocotools` and `pycocoevalcap`:** Used for handling the COCO dataset and evaluating image captioning models using standard COCO metrics.
  GitHub (pycocotools): `https://github.com/cocodataset/cocoapi`
  GitHub (pycocoevalcap): `https://github.com/salmanbura/pycocoevalcap`
- **`pandas`:** Used for analyzing and managing evaluation results.
  Website: `https://pandas.pydata.org/`
- **`tempfile`:** Used for creating temporary files during evaluation steps.

Pre-trained BLIP Models: We accessed the following model weights and configurations via the Hugging Face Model Hub using the `transformers` library:

- Salesforce/blip-image-captioning-base
- Salesforce/blip2-flan-t5-xl-coco
- Salesforce/blip2-opt-2.7b

6. What we learned

6.1. Individual Contributions and Reflections

Deepika Tendulkar:

Working on this project has deepened my practical understanding of multimodal AI systems, particularly how computer vision and natural language processing can be combined to generate coherent image captions. I gained end-to-end experience in building a training pipeline using BLIP and BLIP-2 models, loading COCO dataset annotations, fine-tuning on domain-specific data, and systematically evaluating results with a variety of metrics including BLEU, METEOR, CIDEr, SPICE, ROUGE-L, and CLIPScore. I learned how the scale and architecture of models like OPT-2.7B and Flan-T5-XL affect performance and generalization. Collaborating on the presentation and report writing improved my ability to summarize complex technical insights and communicate results clearly. This project gave me a strong foundation in model evaluation, architecture selection, and understanding the practical challenges of working with large-scale models.

Sanjana Vegesna:

This project has significantly enhanced my understanding of deep learning models, particularly in the context of image captioning. I gained hands-on experience in implementing and fine-tuning BLIP models, which deepened my knowledge of their architecture and capabilities. A key takeaway was a more profound understanding of transformer technology, especially Vision Transformers (ViTs) and their application in processing image data. I learned how these models can be adapted for multi-modal tasks. Managing the dataset, including collection and preprocessing, provided valuable insights into the importance of data handling in deep learning projects. Furthermore, I developed practical skills in model loading, visualization, and inference scripting. Overall, this project has equipped me with a comprehensive skill set in image captioning and a deeper appreciation for the power and versatility of transformer networks.

7. Summary

This project explored and compared multiple image captioning models—BLIP (baseline and fine-tuned), BLIP-2 OPT 2.7B, and BLIP-2 Flan-T5-XL—on the COCO dataset. Through rigorous experimentation, we observed that fine-tuning significantly enhances performance, especially in models like BLIP. However, larger-scale models such as BLIP-2 OPT 2.7B demonstrated the strongest results in metrics like CIDEr and BLEU-4, showcasing the advantage of model size and architecture in capturing complex image-text relationships.

Our evaluation also highlighted the importance of diverse metrics including CLIPScore, which added a semantic layer to performance analysis. While all models showed promising capabilities, the project emphasized that model selection should be guided by application-specific needs—such as speed, interpretability, or resource constraints.

Overall, this work not only reinforced our understanding of transformer-based vision-language models but also offered insights into the practical trade-offs involved in deploying them. The results suggest promising directions for future research, such as fine-tuning large models on domain-specific data, exploring multilingual captioning, or integrating captioning with downstream tasks like visual QA or retrieval.

7.1. References

References

[1] S. Huang, Z. Pan, and Y. Bai. Image Captioning with Transformer-based Models: A Survey. *ACM Computing Surveys*, 56(1):1–35, 2024. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3617592

[2] T.-Y. Lin et al. Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision (ECCV)*, 2014. [Online]. Available: https://cocodataset.org/

[3] A. Gupta. Step-by-Step Guide to Build Image Caption Generator using Deep Learning. *Analytics Vidhya*, Dec. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-c\\aption-generator-using-deep-learning/

[4] TensorFlow. Image Captioning. 2024. [Online]. Available: https://www.tensorflow.org/text/tutorials/image_captioning

[5] Hugging Face. Image Captioning. *Transformers Documentation*, ver. 4.35.1, 2024. [Online]. Available: https://huggingface.co/docs/transformers/v4.35.1/tasks/image_captioning

[6] Hugging Face. BLIP Documentation. 2023. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/blip

[7] Hugging Face. BLIP-2 Documentation. 2023. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/blip-2

[8] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*, 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[9] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping Vision-Language Learning with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[10] J. Hessel, A. Holtzman, M. Forbes, and Y. Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*, 2021. [Online]. Available: https://arxiv.org/abs/2104.08718

[11] M. Inuwa. Vision Transformers (ViT) in Image Captioning Using Pretrained ViT Models. 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/06/vision-transformers/

[12] E. Randellini. Image and text features extraction with BLIP and BLIP-2: How to build a multimodal search engine. 2023. [Online]. Available: https://medium.com/@enrico.randellini/image-and-text-features-extraction-w\ith-blip-and-blip-2-how-to-build-a\-multimodal-search-engine-a4ceabf51fbe