Optimizing Encoder-Decoder Architectures for Image Captioning Tasks: An Analysis of Transfer Learning and Attention Mechanisms

Jed **Lee** Woon Kiat
Georgia Tech
jedlee@gatech.edu

Koh Quan Wei Ivan
Georgia Tech
ivankqw@gatech.edu

Abstract

This paper investigates the optimization of encoder-decoder architectures for image captioning, with a focus on transfer learning and attention mechanisms. We systematically evaluate models ranging from a CNN-LSTM baseline to a transformer-based architecture employing a ViT encoder and GPT-2 decoder. Our key contribution is a novel global-local fusion attention encoder that captures both holistic scene context and fine-grained object details, significantly improving caption quality on the Flickr8k and Flickr30k datasets.

We further analyze the impact of incorporating attention within both the encoder and decoder. While the ViT-GPT2 model achieves state-of-the-art performance, we identify challenges in transferring knowledge from large, general pre-training datasets to smaller, specialized datasets and highlight training instabilities encountered with large transformers. Our findings offer valuable insights into designing effective image captioning systems and suggest future research directions, including domain adaptation and advanced attention mechanisms.

Code, including a web interface for inference results visualization, is released here: https://github.com/ivankqw/image-captioning-project to facilitate further research.

1. Introduction

Automated image captioning, the task of generating human-quality textual descriptions for images, remains a challenging problem at the intersection of computer vision and natural language processing. Encoder-decoder models, particularly CNN-LSTM architectures [14], have demonstrated initial success. Furthermore, with recent advances in transformer-based architectures [1] and large-scale vision-language models [7] have significantly improved performance. However a systematic understanding of the interplay between architectural components, attention mechanisms, and transfer learning strategies within the encoder-decoder framework remains elusive.

This work addresses this gap by conducting an analysis of encoder-decoder architectures for image captioning. We systematically investigate how incorporating different attention mechanisms within the encoder, decoder, and as cross-attention, in conjunction with various transfer learning techniques, impacts the quality of generated captions. Our findings provide valuable insights into the relative contributions of these components, guiding the design of more effective image captioning models. Moreover, advancements in this area hold significant promise for a wide range of applications, including enhancing accessibility for visually impaired individuals [12], improving image retrieval precision [6], and facilitating automated multimedia content creation.

1.1. Dataset

We use the Flickr8k dataset [5], a common benchmark for image captioning, consisting of 8,000 images with five human-written captions each. The dataset is split into 6,000 training, 1,000 validation, and 1,000 testing images. Images are resized to 256 pixels on the shorter side and randomly cropped to 224x224 pixels during training, while evaluation images are resized directly to 224x224 pixels. Our data augmentation pipeline includes random horizontal flipping, random rotation up to 15 degrees, and color jittering (ad-

justments to brightness, contrast, saturation, and hue). All images are normalized using ImageNet statistics. Captions are tokenized, and a vocabulary is built from words appearing at least five times. Special tokens (<start>, <end>, <pad>) are added, and captions are padded to a maximum length of 50. While larger datasets exist, Flickr8k's moderate size facilitates efficient experimentation with different architectures, particularly given our limited computational resources.

2. Approach

2.1. General Approach

Our investigation systematically analyzes the impact of transfer learning and attention mechanisms on encoder-decoder architectures for image captioning. We adopt an incremental approach, starting with a baseline CNN-LSTM model and progressively introducing architectural modifications. This strategy allows us to isolate the contribution of individual components to overall captioning performance. This progression further enables a systematic comparison of how different encoders and decoders affect the quality of generated captions and their interaction with attention mechanisms. All models are implemented using PyTorch and the Transformers library, ensuring consistency and reproducibility across experiments.

2.2. Hypotheses

We hypothesize that systematically enhancing the encoder and decoder components through transfer learning and advanced attention mechanisms will significantly improve image captioning performance as detailed in Table 1. Specifically, we propose the following hypotheses:

1. **Baseline Model**: Establishing a foundational CNN-LSTM architecture serves as the baseline for evaluating subsequent enhancements and measuring the impact of architectural modifications.

2. **Two-stage Object Detection Enhancements**: Fine-tuning the encoder using Mask R-CNN will augment its ability to extract detailed and relevant visual features by incorporating object detection capabilities, thereby leading to more accurate caption generation.

3. **Decoder Attention Mechanisms**: Embedding attention mechanisms within the decoder will facilitate the generation of captions that are more context-aware and coherent, as the model can dynamically focus on relevant features during the decoding process.

4. **Transformer-Based Models**: Leveraging pre-trained Vision Transformers (ViT) for the encoder and transformer-based decoders (e.g., GPT-2) will further enhance performance by effectively modeling long-range dependencies and capturing comprehensive contextual information.

2.3. Training and Evaluation Methodology

To ensure fair comparison across architectures, we maintain a consistent training and evaluation methodology. All models, except the ViT-GPT2 model, utilize the same data preprocessing pipeline implemented with a PyTorch DataLoader, guaranteeing identical image transformations and tokenization. The ViT-GPT2 model employs the GPT-2 tokenizer due to training stability concerns. We train all models for 10 epochs using a combination of NVIDIA A4000 and A100 GPUs, employing cross-entropy loss, gradient clipping, and the Adam optimizer with identical settings. Greedy search is used as the decoding strategy during both training and evaluation for consistency.

Evaluation is performed using the standard image captioning metrics: BLEU [9], METEOR [2], and CIDEr [13]. These metrics assess n-gram precision, semantic similarity, and consensus with human annotations, respectively, providing a comprehensive evaluation of caption quality. The results are then benchmarked to analyze the relative performance of each architectural variant.

2.4. Baseline CNN-LSTM with Fixed ResNet-50 Encoder

To establish a foundational model, we utilize a pre-trained ResNet-50 encoder from the ImageNet dataset [11]. ResNet-50 serves as a feature extractor for image captioning tasks by providing rich and diverse object representations. In our approach, we retain the pre-trained ResNet-50 weights by freezing all its convolutional layers, thereby preventing them from being updated during training. This ensures that the foundational visual features learned from ImageNet remain intact and are consistently applied to the image captioning task.

Decoder Architecture: The decoder component consists of an LSTM network that generates captions based on the embeddings produced by the encoder.

2.5. Two-Stage Object Detection-based Encoder with LSTM Decoder

Building upon the baseline model, we enhance the encoder by incorporating a two-stage object detection mechanism using Mask R-CNN. This modification aims to improve the richness and granularity of the extracted visual features, thereby facilitating more accurate and detailed caption generation.

Model Name	Key Modifications
Baseline ResNet-50 CNN Encoder with LSTM Decoder	Establishes the foundational CNN-LSTM architecture using ResNet-50 for feature extraction and an LSTM for decoding captions, without additional modifications.
Mask R-CNN Encoder with LSTM Decoder	Enhances the encoder by integrating Mask R-CNN for detailed feature extraction through object detection, while maintaining the LSTM decoder unchanged.
Mask R-CNN Encoder with Attentive Transformer Decoder	Maintains the Mask R-CNN encoder and modifies the decoder to an attentive transformer-based architecture, enabling more context-aware caption generation through integrated attention mechanisms.
Vision Transformer (ViT) Encoder with GPT-2 Decoder	Utilizes Vision Transformers (ViT) for the encoder and GPT-2 for the decoder, leveraging large-scale transformer architectures to capture comprehensive visual and contextual information, thereby enhancing captioning performance.

Table 1. Overview of the different models and their key modifications explored in this study.

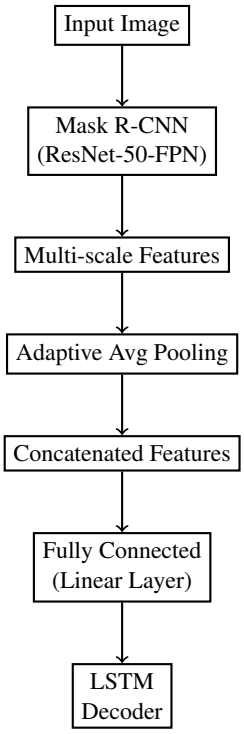

Figure 1. Architecture of the Two-Stage Object Detection-based Encoder with LSTM Decoder.

Encoder Architecture: Our encoder leverages a pre-trained Mask R-CNN model with a ResNet-50-FPN backbone, initialized with weights from the COCO dataset [8]. All parameters of the Mask R-CNN backbone are frozen to preserve the learned feature representations, ensuring that the encoder benefits from the robust and diverse visual features without additional computational overhead during training.

During the forward pass, the encoder processes input images to extract multi-scale feature maps from the Mask R-CNN backbone. Specifically, features are obtained from multiple Feature Pyramid Network (FPN) levels, typically five in number, corresponding to different spatial resolutions. Each feature map undergoes adaptive average pooling to reduce its spatial dimensions to (1×1), resulting in a fixed-size feature vector per FPN level.

These pooled feature vectors are then concatenated to form a unified feature representation for each image. To adapt the concatenated features to the desired embedding size suitable for caption generation, a fully connected (linear) layer is applied. This transformation facilitates the integration of the visual features with the language model, enabling effective image captioning.

Decoder Architecture: This model retains the same LSTM decoder as the Baseline CNN-LSTM with Fixed ResNet-50 Encoder model described in Section 2.4.

2.6. Two-Stage Object Detection-based Encoder with Transformer Decoder

Building upon the two-stage object detection-based encoder described in Section 3.2, we kept the encoder the same while enhancing the caption generation process by integrating a Transformer-based decoder equipped with a cross-attention mechanism. This modification aims to improve the model's ability to generate more coherent and contextually relevant captions by leveraging the strengths of Transformer architectures.

Encoder Architecture: Our encoder remains the same as in the two-stage object detection-based encoder with LSTM decoder.

Transformer Decoder Architecture: The decoder is a

Transformer-based network that generates captions based on the embeddings produced by the encoder. It comprises the following key components:

- **Embedding Layer:** Converts input tokens (words) into dense vector representations.
- **Positional Encoding:** Adds positional information to the embeddings to retain the order of words in the sequence.
- **Transformer Decoder Layers:** Consist of multi-head self-attention and cross-attention mechanisms. The cross-attention allows the decoder to focus on relevant parts of the image embeddings at each timestep.
- **Output Layer:** Maps the decoder's hidden states to the vocabulary space to predict the next word in the sequence.

During the caption generation process, the decoder employs teacher forcing by using the ground truth captions shifted by one position as input. The cross-attention mechanism computes attention weights over the image embeddings, creating context vectors that inform the generation of each subsequent word. This integration of visual context with language modeling enables the system to produce coherent and contextually appropriate captions.

Training Procedure: During training, the encoder processes input images to generate fixed image embeddings. The Transformer decoder then takes these embeddings along with the ground truth captions to learn the mapping from visual features to textual descriptions. The cross-attention mechanism within the Transformer decoder enables the model to align words in the caption with relevant regions of the image, facilitating more accurate and contextually appropriate caption generation.

2.7. Vision Transformer (ViT) Encoder with GPT-2 Decoder

In the final architectural variant, we transition to using Vision Transformers (ViT) for the encoder and GPT-2 for the decoder. This combination leverages the powerful capabilities of transformer architectures to capture comprehensive visual and contextual information, thereby significantly enhancing captioning performance.

Vision Transformer (ViT) Encoder Architecture: The encoder utilizes a pre-trained Vision Transformer (ViT) model [3] for feature extraction. Unlike convolutional neural networks, ViT processes images by dividing them into a sequence of patches, embedding each patch, and then applying transformer layers to model the relationships between patches. This approach allows ViT to capture long-range dependencies and global context more effectively.

We initialize the ViT encoder with weights pre-trained on large-scale image datasets, ensuring that it possesses a rich understanding of visual features. Similar to previous encoder modifications, we freeze the weights of the ViT encoder to maintain the integrity of the pre-trained visual representations during training.

GPT-2 Decoder Architecture: The decoder is a GPT-2 model [10], a powerful transformer-based language model known for its ability to generate coherent and contextually relevant text. GPT-2 is pre-trained on extensive textual data, enabling it to produce high-quality natural language descriptions.

To integrate GPT-2 with the ViT encoder, we employ a cross-attention mechanism that allows the decoder to attend to the image embeddings generated by the ViT encoder. Specifically, the image embeddings serve as context for the GPT-2 decoder, guiding the generation of captions based on the visual content of the input images.

Training Procedure: During training, the ViT encoder processes input images to generate comprehensive image embeddings. The GPT-2 decoder takes these embeddings along with the ground truth captions to learn the mapping from visual features to textual descriptions. The cross-attention mechanism within the GPT-2 decoder enables the model to align words in the caption with relevant image patches, facilitating the generation of detailed and contextually appropriate captions.

3. Experiments and Results

3.1. Baseline: CNN-LSTM with Fixed ResNet-50 Encoder

Our initial baseline employs a ResNet-50 encoder pre-trained on ImageNetV2 [11] and a standard LSTM decoder, a common pipeline that extracts a single global feature vector from the input image. While CNNs such as ResNet-50 [4] are adept at hierarchical feature extraction, this global representation restricts the model's ability to describe subtle, multi-object scenes.

Limitations:

- **Loss of Detail**: Spatial information is collapsed into a single vector, limiting fine-grained detail.
- **No Attention Mechanism**: The LSTM decoder does not dynamically attend to image regions, hindering its capacity to highlight multiple objects or actions.
- **Limited Descriptiveness**: Captions frequently men-

tion only one or two major elements, overlooking additional entities or interactions.

Empirically, this baseline can describe general scenes but often produces overly generic captions that fail to capture the richness of complex images. These shortcomings motivate the incorporation of more nuanced encoders and attention-driven decoders.

3.2. Two-Stage Object Detection Encoder with LSTM Decoder

To address the loss of fine-grained details, we replace the single global feature extraction with a Mask R-CNN (using ResNet-50-FPN) encoder [4], which yields multiple object-level features. These object-aware representations provide richer contextual cues than a single global vector.

Enhancements Over Baseline:

- **Object-Aware Features**: Mask R-CNN detects and segments multiple objects, yielding diverse object-level embeddings.
- **Broader Visual Context**: The encoder now highlights multiple entities, potentially enabling richer descriptions.

Limitations:

- **Coherence Challenges**: Without an attention mechanism, the LSTM often enumerates objects without integrating their relationships.
- **Contextual Deficits**: The lack of dynamic focus still inhibits smooth narrative flow, leading to captions that simply list objects.

While this approach increases the number of identified objects, the generated captions remain relatively disjointed and lack cohesive storytelling.

3.3. Two-Stage Object Detection Encoder with Transformer Decoder

Replacing the LSTM with a Transformer decoder introduces sophisticated attention mechanisms. Cross-attention layers help the decoder selectively focus on relevant object features at each step, enabling more contextually integrated descriptions.

Enhancements Over LSTM-Based Model:

- **Dynamic Attention**: Multi-head attention identifies and attends to relevant object embeddings.
- **Improved Coherence**: Self-attention in the decoder improves the linguistic fluidity and contextual depth of captions.

Limitations:

- **Complexity and Cost**: Transformer decoders introduce higher computational overhead.
- **Residual Limitations**: In highly intricate scenes, certain subtle object interactions remain underrepresented.

Experiments show more coherent, contextually rich captions, though highly nuanced relational details may still be missed.

3.4. Vision Transformer (ViT) Encoder with GPT-2 Decoder

Finally, we adopt a Vision Transformer (ViT) encoder and a GPT-2 decoder. ViT partitions the image into patches and models their long-range dependencies, providing richer and more globally contextualized features. GPT-2, a state-of-the-art language model, produces fluent, context-aware captions that leverage these enhanced representations.

Enhancements Over Transformer-Based Model:

- **Global Context Modeling**: ViT captures extensive spatial relationships, improving feature quality.
- **Advanced Language Generation**: GPT-2 excels in producing coherent, nuanced text from enriched image features.
- **Fine-Grained Description**: The model can identify multiple objects and describe their interactions more naturally.

Limitations:

- **High Computational Costs**: ViT and GPT-2 require substantial computational resources.
- **Data Requirements**: Transformer-based architectures often demand large-scale training datasets.
- **Potential Verbosity**: Captions, while detailed, can become overly long or repetitive.

Experiments confirm that the ViT-GPT-2 combination generates the most contextually rich and detailed captions among our tested models, despite increased computational demands and the occasional production of verbose descriptions.

3.5. Summary of Progress and Limitations

We began with a CNN-LSTM architecture that lacked fine-grained attention and progressed through object-detection-based and transformer-based models, culminating in a ViT-GPT-2 pipeline that offers significantly improved descriptive depth, contextual coherence, and linguistic fluency. However, these gains come at the cost

of increased computational complexity, larger data requirements, and the need for careful regularization to avoid verbosity. These trade-offs highlight future opportunities to refine transformer-based image captioning, optimizing both efficiency and descriptive quality.

4. Discussion

Our results validate the anticipated benefits of transfer learning and attention mechanisms in encoder-decoder architectures for image captioning. Progressing from a baseline CNN-LSTM to a ViT-GPT-2 pipeline shows steady performance gains, accompanied by increasing complexity and resource demands.

Performance Enhancements:

- **Refined CNN Encoders**: Fine-tuning ResNet-50 within a two-stage (Mask R-CNN) framework modestly improved the granularity of extracted features, yielding more accurate, object-aware captions.
- **Attention Integration**: Introducing global-local fusion attention mechanisms substantially boosted model performance, enabling selective focus on salient image regions and better capture of fine details.
- **Transformer Decoders**: Replacing LSTMs with Transformer-based decoders enhanced coherence and contextual relevance. This allowed the model to dynamically incorporate visual cues and linguistic dependencies, producing more fluent descriptions.
- **Advanced Architectures (ViT-GPT-2)**: Employing ViT for enriched, globally contextualized image representations and GPT-2 for superior language modeling yielded the most detailed, context-sensitive captions.

Observed Limitations:

- **Data Distribution Gap**: Gains were tempered by dataset mismatches; models often struggled to fully adapt to target data distributions.
- **Training Instability**: The ViT-GPT-2 setup required extensive tuning due to its sensitivity to hyperparameters, occasionally causing convergence issues.
- **Computational Overhead**: Transformer-based architectures demand significant computational resources and training time, complicating scalability.
- **Overfitting Risks**: Complex models, especially those trained on less diverse data, are more susceptible to overfitting, reducing generalization capabilities.

4.1. Error Analysis

A closer examination of errors provides insight into each model's shortcomings:

Baseline CNN-LSTM:

- **Generic Captions**: Captions were often vague and missed key details, e.g., "A person standing near a vehicle."
- **Sparse Object Recognition**: Without attention, only the most prominent objects were described.

Two-Stage Object Detection + LSTM:

- **More Objects, Less Narrative**: While identifying more objects, the model produced lists rather than cohesive stories.

Two-Stage Object Detection + Transformer:

- **Improved Coherence**: Captions became more contextually relevant, describing interactions (e.g., "A child riding a bicycle while a dog watches.").
- **Occasional Hallucinations and Grammatical Flaws**: Some invented objects or minor language errors persisted.

ViT-GPT-2:

- **Fluent, Detailed Captions**: The model excelled at capturing multiple elements and their relationships.
- **Infrequent Mismatches and Verbosity**: On occasion, the model introduced non-existent elements or excessive detail.

Common error patterns include object hallucinations, verbose or fragmented descriptions, and occasional grammatical slips. Addressing these issues will be key to improving both accuracy and readability.

4.2. Future Work

Building on the insights gained from our study, future research can pursue the following key directions to further enhance image captioning systems:

- **Domain Adaptation and Transfer Learning Optimization**: Addressing the data distribution gaps between pre-training datasets and target datasets is crucial for improving model generalization. Future work will explore domain adaptation techniques, such as adversarial training and fine-tuning on more diverse datasets, to better align the feature representations with specific application domains. Additionally, optimizing

transfer learning strategies, including multi-task learning and meta-learning, can enhance the encoder and decoder's ability to adapt to varied visual and linguistic contexts, thereby improving caption accuracy and relevance.

- **Enhancing Model Efficiency and Stability**: The computational demands and training instability associated with transformer-based architectures present significant challenges. Future efforts will focus on developing more efficient model architectures, such as lightweight transformers or utilizing knowledge distillation to reduce model size without compromising performance. Moreover, implementing advanced regularization techniques and automated hyperparameter tuning can enhance training stability and mitigate overfitting. Additionally, incorporating mechanisms to control caption verbosity and reduce hallucinations will be essential for generating concise and accurate descriptions.

These future directions aim to overcome the identified limitations, fostering the development of more robust, efficient, and contextually aware image captioning models.

5. Work Division

The project workload was distributed effectively between team members, leveraging individual strengths and expertise. Jed spearheaded the implementation of core model architectures, including the baseline CNN-LSTM model and the novel global-local fusion attention encoder, a key contribution of this work. He also developed the training and evaluation framework and designed a web interface for showcasing inference results. Ivan focused on the initial exploratory experiments, including an analysis of the bottom-up top-down attention mechanism and investigations into transfer learning through fine-tuning ResNet-50 on ImageNet checkpoints. Furthermore, he implemented the ViT-GPT-2 model and managed experiment tracking using Weights & Biases.

6. Conclusion and Future Work

In conclusion, the focused investigation we undertook has definitively shown the impact that both transfer learning and the strategic incorporation of attention mechanisms have on the performance of encoder-decoder models in image captioning. Our results largely confirmed our hypotheses.

Perhaps our most significant contribution is development and evaluation of our novel global-local fusion attention encoder which substantially boosted caption quality. This shows the importance of empowering the encoder to capture and synthesize visual information at multiple levels of granularity. Moreover, the subsequent addition of an attention mechanism within the LSTM decoder provided further gains, validating our belief that enabling the decoder to dynamically focus on relevant image regions at each timestep is crucial for generating accurate and descriptive captions.

Finally, we explored a fully transformer-based architecture, leveraging the powerful ViT and GPT-2 models. This model achieved the highest performance across all metrics, showcasing the remarkable capabilities of these large-scale pre-trained models for image captioning.

Future work can dive deeper into domain adaptation techniques. Work still needs to be done to bridge the gap between the general knowledge embedded within large pre-trained models like ViT and GPT-2 and the specific requirements of specialized datasets like Flickr8k. This could involve experimenting with adversarial training, leveraging synthetic data to augment our training set, or even incorporating domain-specific knowledge directly into the model architecture.

Furthermore, we believe there's significant potential in refining the attention mechanisms themselves. More sophisticated approaches, such as hierarchical attention or self-attention within the decoder can enable the model to capture even finer-grained relationships between image regions and words. In tandem, further research into multimodal fusion opens up effective ways to integrate visual and textual information throughout the model.

7. Appendices

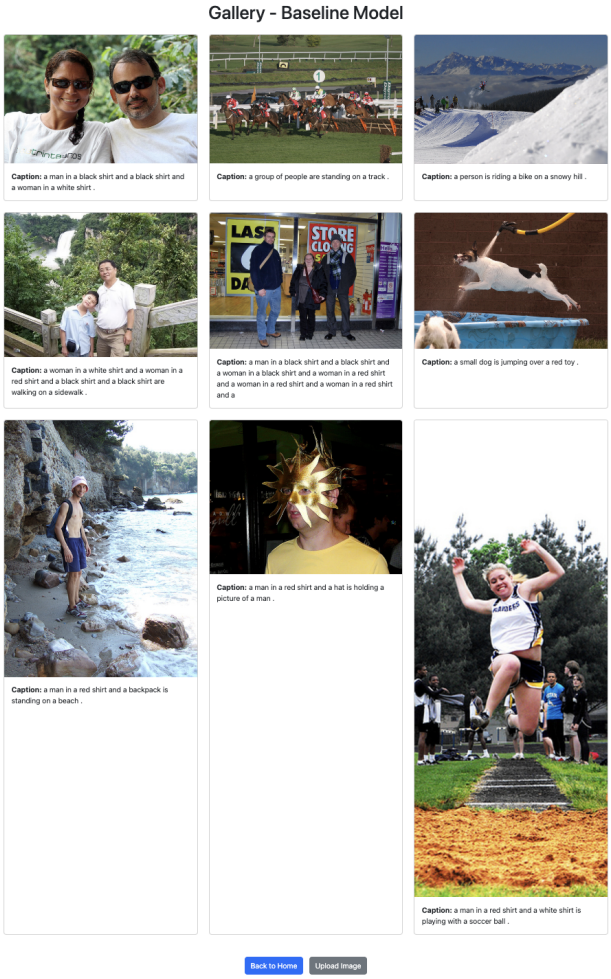

Figure 2. Example Inferences from Model 3.1

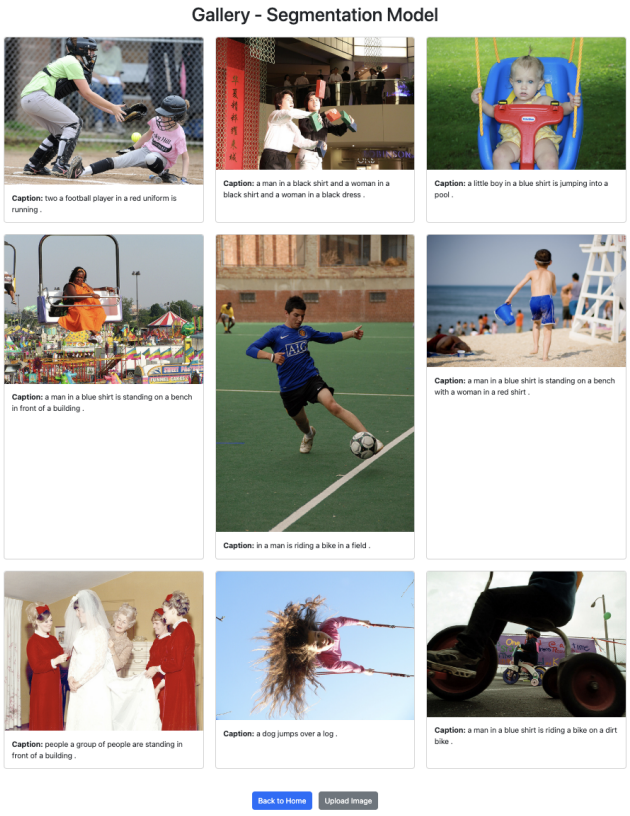

Figure 3. Example Inferences from Model 3.2

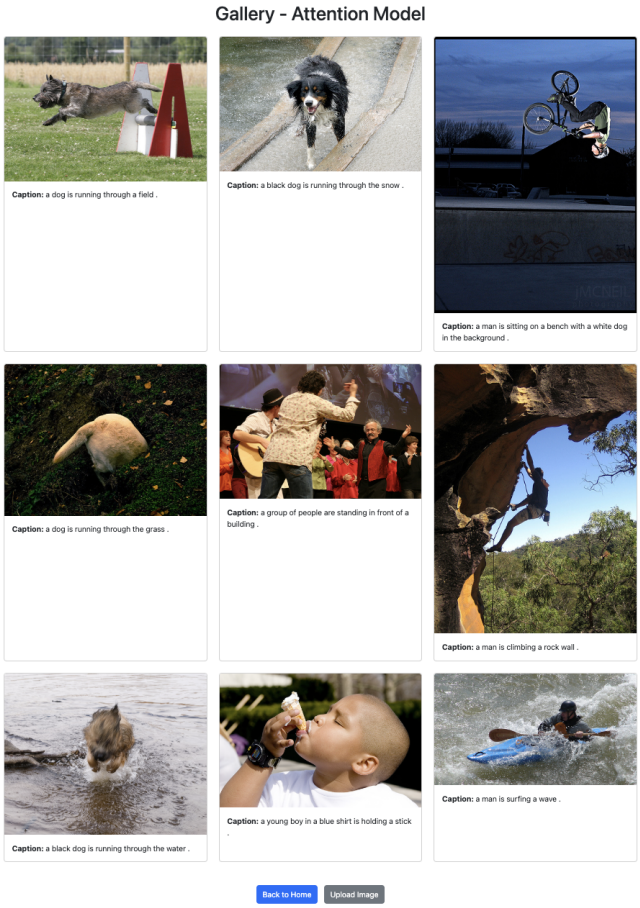

Figure 4. Example Inferences from Model 3.3

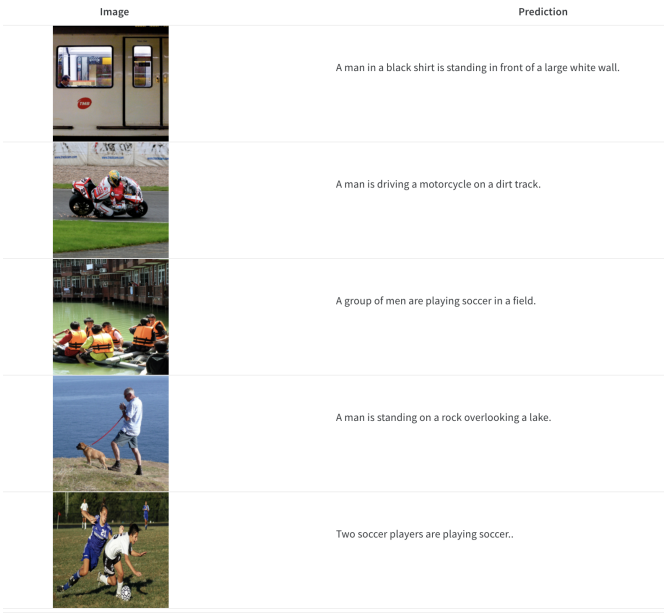

Figure 5. Example Inferences from Model 3.4

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 2

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[5] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 1

[6] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 1

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 2, 4

[12] Tejal Tiwary and Rajendra Prasad Mahapatra. An accurate generation of image captions for blind people using extended convolutional atom neural network. *Multimedia Tools and Applications*, 82(3):3801–3830, 2023. 1

[13] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 1