1. INTRODUCTION

Image captioning is the task of automatically generating natural language descriptions for images. It lies at the intersection of computer vision and natural language processing and has broad applications in areas such as assistive technologies for the visually impaired, content moderation, image retrieval, and human-computer interaction. Despite recent progress in deep learning, generating captions that are both accurate and coherent remains a challenging problem, especially when attention to relevant image regions is crucial for context-aware descriptions.

This project addresses the image captioning problem using a deep learning-based encoder-decoder architecture with attention mechanisms. Specifically, a pretrained convolutional neural network (CNN) encodes visual features from an input image, which are then decoded into a caption using a recurrent neural network (RNN) enhanced with an attention mechanism. The attention component allows the model to focus on specific parts of the image while generating each word, improving both interpretability and performance.

The dataset used, MS COCO, contains thousands of images each paired with 5 human-generated captions. This data provides a rich source of visual and linguistic knowledge for training models that learn to "translate" visual content into language.

This project not only demonstrates practical application of attention-based deep learning architectures but also contributes to the broader goal of making AI systems more interpretable and accessible. By visualizing attention maps, users can gain insight into the model's decision-making process — a step toward explainable AI in vision-language tasks.

2. STATE-OF-THE-ART

Image captioning has evolved from simple template-based systems to deep learning models capable of generating fluent and context-aware descriptions. A foundational approach involves CNN-RNN architectures, where a convolutional neural network encodes the image and a recurrent network, often an LSTM, decodes it into a sentence [2]. These models are intuitive and effective but suffer from compressing all visual information into a fixed-length vector, limiting their ability to generate detailed or varied descriptions, especially with complex images .

To overcome this, attention mechanisms were introduced, allowing the model to dynamically focus on different image regions while generating each word [1]. This not only improved caption quality but also provided interpretability by aligning visual features with words. However, RNNs still posed limitations in capturing long-range dependencies and parallelization. Transformer-based models addressed these issues, using self-attention to process image and text jointly [3]. Though these newer models (e.g., ViLBERT, BLIP) achieve state-of-the-art results, they require large datasets and hardware, and are often less transparent.

In our project, we deliberately chose to build on the attention-augmented CNN-RNN architecture due to its transparency, lower complexity, and educational value. Starting from an open-source implementation of the standard encoder-decoder model [4], we extended it by integrating soft attention and top-p sampling, and developed a user-facing interface to visualize attention maps aligned with each word in the caption. This design places our work within a classic yet still relevant family of models, while focusing on enhancing interactivity and understanding rather than chasing cutting-edge benchmarks.

3. METHODOLOGY: DATA ANALYSIS

We used the MS COCO dataset, which contains over 80,000 images with five human-written captions each. These captions vary in length, vocabulary, and complexity, providing a rich linguistic signal for learning grounded representations. As part of preprocessing, we **randomly sample one caption per image** per epoch. This reduces memory usage and training time while still providing diverse supervision across epochs.

Data exploration reveals that most captions are short, typically between 5–15 words, but some outliers contain more than 30 words (Figure 1). Additionally, we visualized the frequency of words across the dataset (Figure 2), which highlights a **strong vocabulary imbalance**: a small number of stopwords like "a", "the", and "on" dominate the corpus. To mitigate this and reduce noise from rare words, we exclude words that appear fewer than four times from the vocabulary. This balances expressiveness with generalization, resulting in a vocabulary of around 9,000 words.

On the image side, MS COCO exhibits significant **size variation** (e.g., 480×640, 640×480, etc.). To ensure uniformity and compatibility with pretrained CNN backbones, we resize all images to 256×256 pixels, then apply random cropping to 224×224 and horizontal flipping. These **data augmentations** introduce view diversity, which improves robustness and reduces overfitting.

All images are normalized using the standard ImageNet mean and standard deviation, ensuring alignment with pre-trained feature extractors. Captions are lowercased, tokenized, and padded to the maximum length in each batch. Padding tokens are masked during loss computation to prevent them from influencing training.

By identifying **key data characteristics** like caption length distribution, vocabulary frequency, and image size variability — and addressing outliers and imbalances through filtering and normalization — we ensure that the model sees diverse, balanced, and well-prepared inputs. These steps are essential to enabling efficient learning and strong generalization in image captioning.

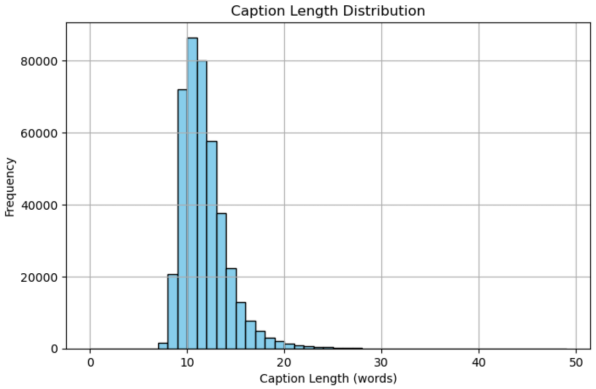

Figure 1: Caption length distribution

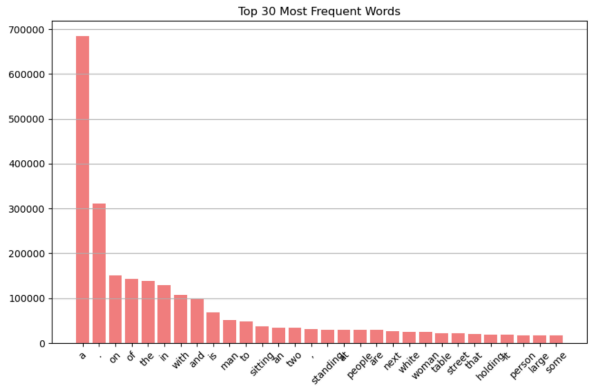

Figure 2: Frequency of words across the dataset

4. METHODOLOGY: MODEL AND OPTIMIZATION

This section details the model architecture, training procedure, and optimization strategies used. Due to limited computational resources and time, we were unable to perform extensive hyperparameter tuning. Instead, we adopted well-established settings from a reference implementation that had been shown to work reliably under similar configurations.

Our image captioning model is built on an encoder-decoder architecture with an attention mechanism, a proven structure in vision-to-language tasks. The encoder is based on a ResNet-152 convolutional neural network pre-trained on ImageNet. We remove the final fully connected and pooling layers and retain the convolutional blocks to extract spatial feature maps from input images. An adaptive average pooling layer downsamples the output to a fixed spatial size of 14×14, resulting in feature tensors of shape (batch_size, 14, 14, 2048). To reduce overfitting and leverage the robustness of pre-trained features, most of the ResNet layers are frozen; only the layers from stage 3 onward are fine-tuned during training.

We incorporate an additive attention mechanism, similar to the one introduced by Bahdanau et al., to allow the decoder to dynamically focus on different regions of the image at each decoding step. The attention weights are computed from both the encoder's output and the decoder’s current hidden state, and they are used to calculate a weighted sum over the image features, forming a context vector that guides word generation. This mechanism helps the model align the generated words with relevant visual content.

The decoder is a single-layer LSTMCell with a hidden size of 512. At each time step, it takes as input the context vector produced by the attention module and the embedded representation of the previously generated word. It updates its hidden state and uses it to predict the next word via a fully connected output layer. A learned gating mechanism (f_beta) modulates the influence of visual features during decoding, and a dropout layer with probability 0.5 is applied to mitigate overfitting.

Training is performed using the Cross Entropy Loss, computed between the predicted word distributions and the ground truth tokens. Padding tokens are ignored in the loss calculation by setting the `ignore_index` parameter, ensuring the model is not penalized for padded elements in variable-length sequences—a standard practice in sequence generation tasks. In addition to cross-entropy, we also tracked perplexity during training, as it provides an

interpretable measure of the model's uncertainty when generating tokens. Though cross-entropy is typically the primary loss function, perplexity is a useful derived metric for monitoring progress and comparing model snapshots over time.

For optimization, we employ the Adam optimizer with a fixed learning rate of 0.001, chosen for its efficiency and stability in training deep networks. Due to time and resource limitations, we did not implement learning rate scheduling, gradient clipping, or other regularization strategies, though these represent viable directions for future work. All decoder parameters are trainable, while the encoder remains partially frozen to preserve learned visual representations and reduce the computational load.

During inference, we use top-p (nucleus) sampling with p = 0.95 instead of greedy decoding. This method selects the next word from a dynamically determined subset of the vocabulary that accounts for 95% of the probability mass, balancing fluency with diversity. It often results in more varied and natural-sounding captions compared to deterministic strategies.

We were constrained from conducting full hyperparameter tuning due to the computational demands of training on the full COCO dataset. Each training run required several hours (even days), making systematic exploration of hyperparameters infeasible. As such, we used the same values as the github implementation we started from, which included a batch size of 128, an embedding size of 256, a hidden size of 512, and a learning rate of 0.001. We trained for 5 epochs, a number that proved effective as the training and validation loss curves showed clear signs of stabilization by the final epoch, indicating the model had sufficiently converged without signs of overfitting. While more rigorous tuning would likely yield performance improvements, the consistency of the results suggests these choices were a sound baseline.

5. EXPERIMENTS

We began our experimentation by evaluating the pretrained CNN-RNN model provided in the original GitHub repository. While this model was capable of generating grammatically correct captions, we quickly discovered that it produced the same caption every time for the same input image, which seemed unnatural since we, as humans, don't usually describe the same thing the same way every time, we introduce some natural variability. Moreover, the model's outputs were not interpretable—there was no way to understand which parts of the image influenced each predicted word, making its decisions opaque and difficult to analyze.

Motivated by these shortcomings, we redesigned the decoder to incorporate a visual attention mechanism. Our hypothesis was that attention would help the model align words with specific image regions, improving performance and offering insight into the model's decision-making through attention maps. In parallel, we replaced the deterministic argmax decoding strategy with top-p (nucleus) sampling to encourage more diverse and context-aware captions. The goal of these architectural and decoding changes was to move away from rigid, repetitive outputs toward more fluent and image-specific descriptions.

Initially, we trained the modified model on a small subset of 1,500 images (using the same setup as the final model, described earlier), but this led to severe overfitting and repetitive, low-quality captions. We attempted to mitigate this by gradually scaling the dataset size to

4,000 and then 8,000 images. While some improvement in variability was observed, the captions remained shallow and repetitive, indicating persistent underfitting in terms of data and overfitting in terms of model capacity. Even scaling further and training on 25,000 images showed limitations due to insufficient scale and compute power. All these attempts were performed using ResNet-50 as the pretrained encoder, also for efficiency reasons.

At this point, we developed a simple user interface (UI) to streamline testing and improve interpretability. The UI allowed us to upload any image—whether from the validation set or external sources—and generate a caption along with attention maps showing which regions the model focused on for each word. This tool helped us qualitatively evaluate the model's behavior, especially the impact of attention on the final captions.

A major breakthrough came when we gained access to a GPU-enabled machine, allowing us to train the model on the full MS COCO training set of over 80,000 images. This marked a turning point in performance: caption quality improved significantly in terms of fluency, diversity, and alignment with the image content. Unlike the earlier subset experiments, the model began to generalize better across unseen images, producing more descriptive and relevant captions. The availability of a full dataset allowed the attention mechanism and top-p sampling to be fully leveraged, as the model had sufficient examples to learn meaningful image-language associations.

To further test the impact of encoder quality, we upgraded the backbone from ResNet-50 to ResNet-152. Our assumption was that deeper image features would enhance the richness of the visual representation and enable more precise grounding of words in image regions. Despite an increase in training time, this change yielded measurable improvements: the model became more adept at recognizing fine-grained objects and constructing more contextually appropriate captions. This upgrade, combined with attention and top-p sampling, allowed the model to produce outputs that were not only accurate but also interpretable and diverse. Naturally, not all generated captions were perfect — the model occasionally produced incoherent or irrelevant descriptions, especially in complex or cluttered scenes. This is expected given the limitations of the training data and the fact that the model does not perform explicit object detection or reasoning, but instead relies on learned statistical patterns in the data.

6. RESULTS

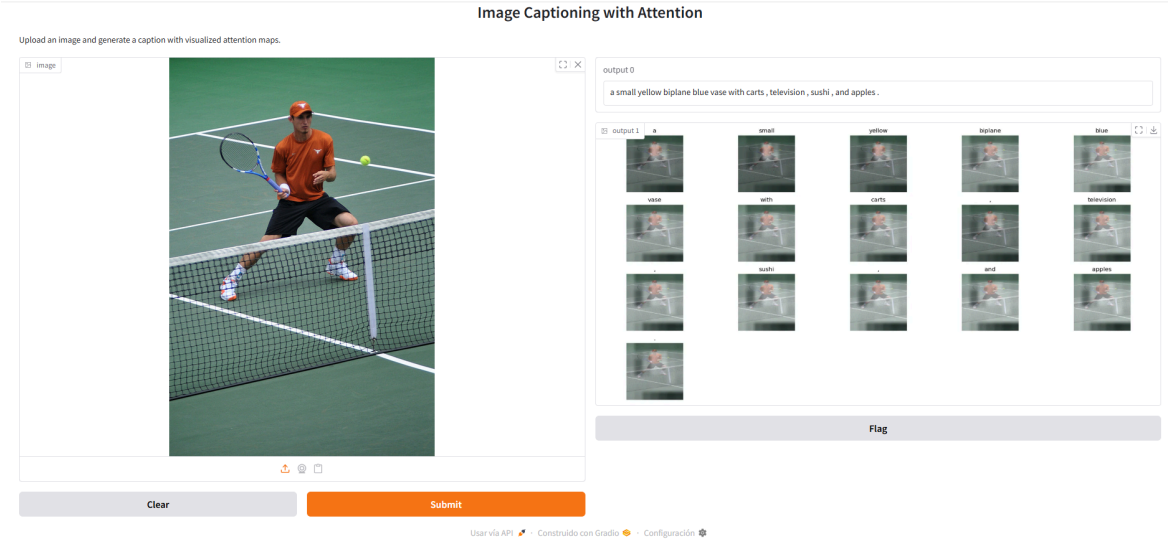

Figure 3: Caption and attention maps of an image generated with model trained with 8k samples

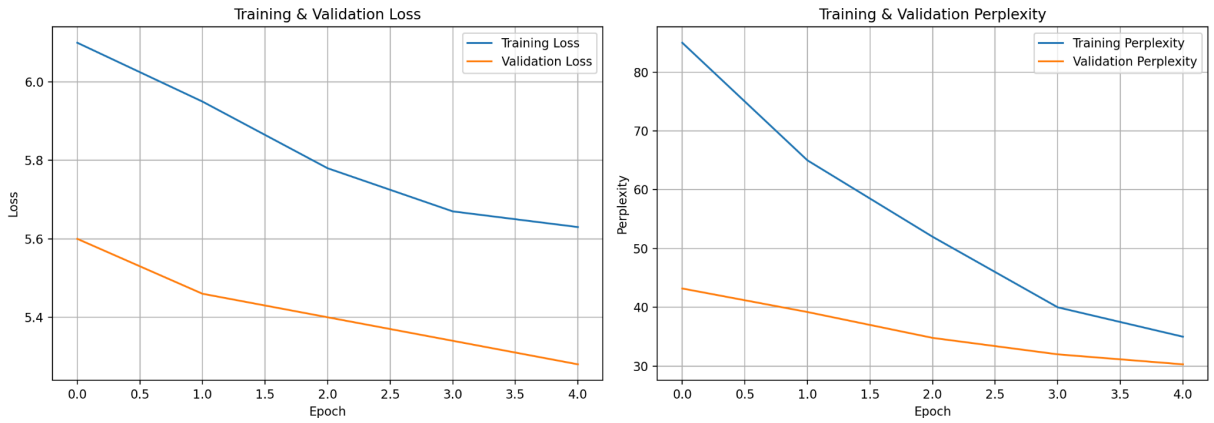

Figure 4: Train and validation loss and perplexity for training with 8k samples

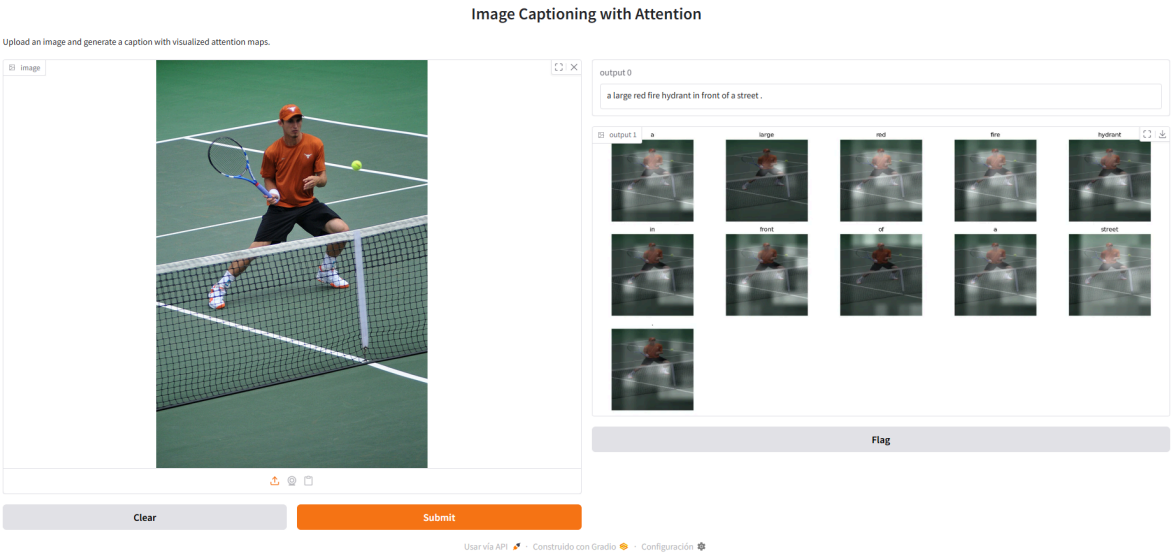

Figure 5: Caption and attention maps of an image generated with model trained with 25k samples

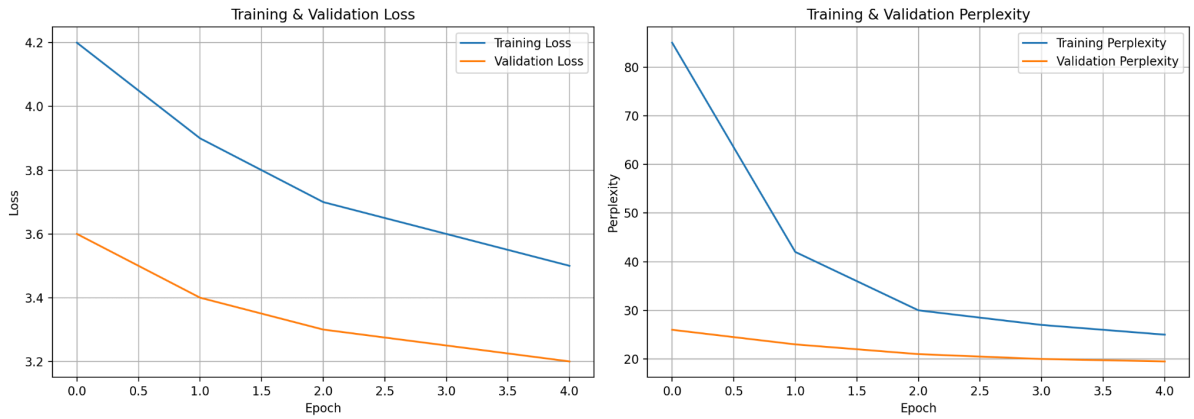

Figure 6: Train and Validation loss and perplexity for training with 25k samples

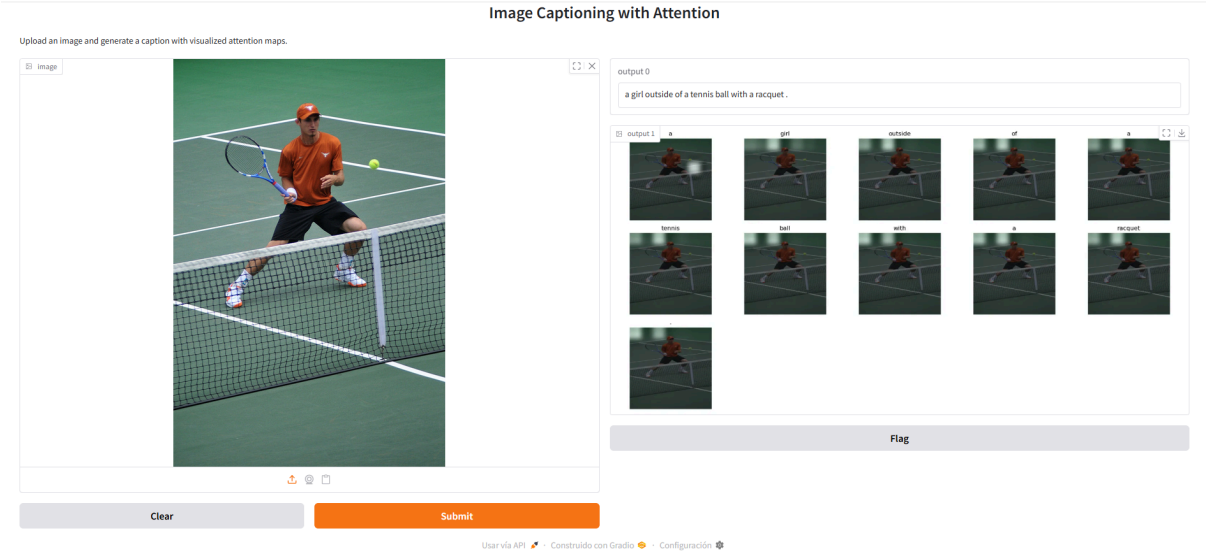

Figure 7: Caption and attention maps of an image generated with model trained with all samples and resnet-50

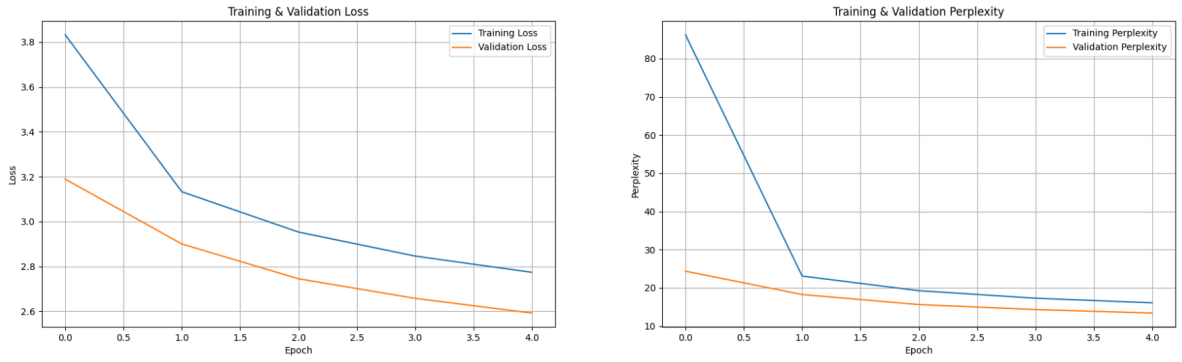

Figure 8: Train and validation loss and perplexity for training with all samples and resnet-50

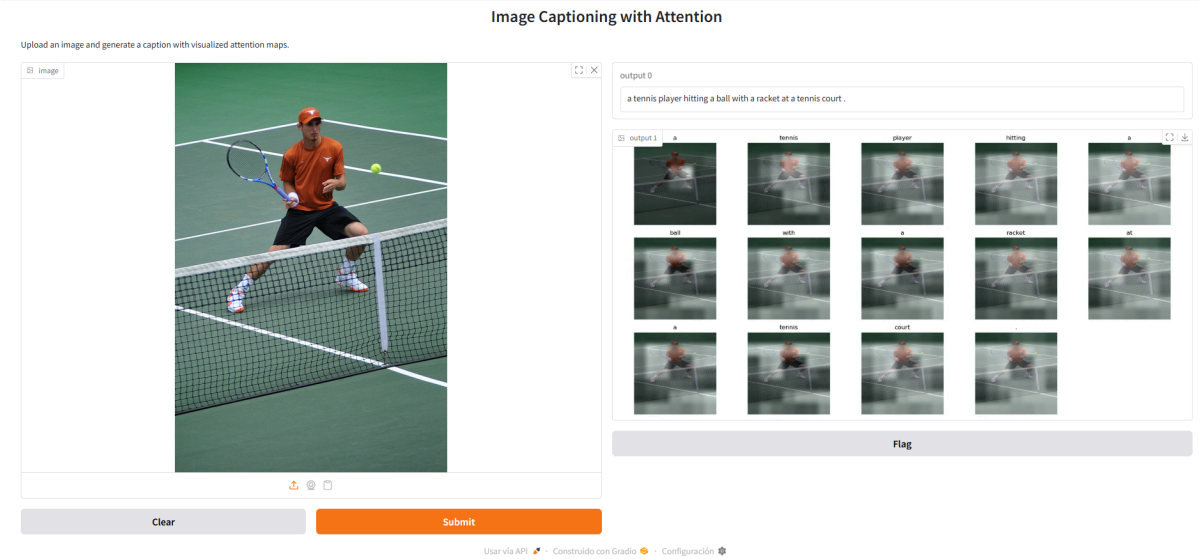

Figure 9: Caption and attention maps of an image generated with model trained with all samples and resnet-152

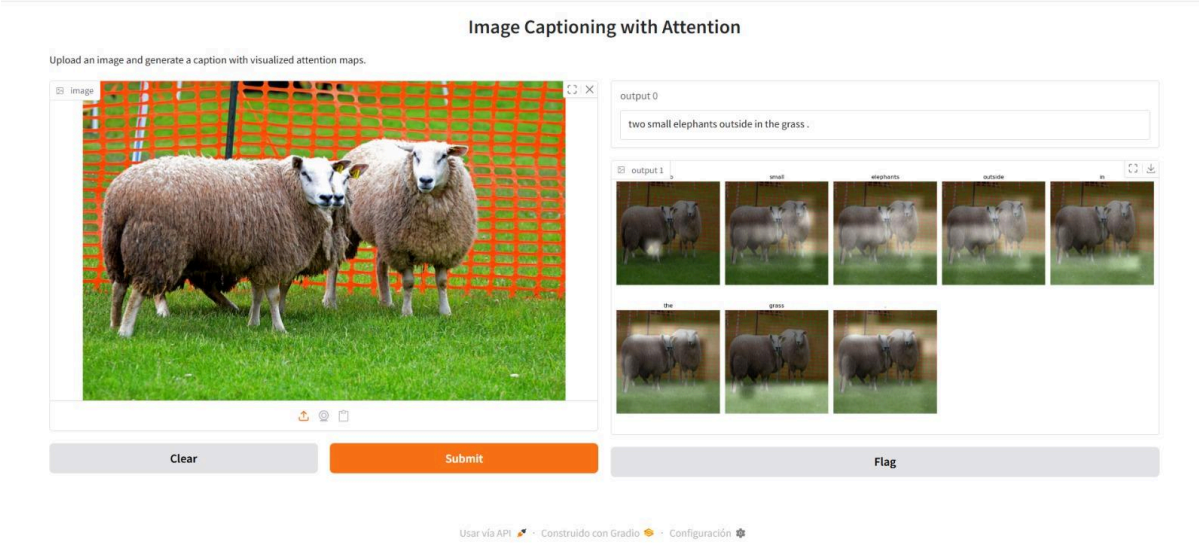

Figure 10: Caption and attention maps of image generated with model trained with all samples and resnet-152

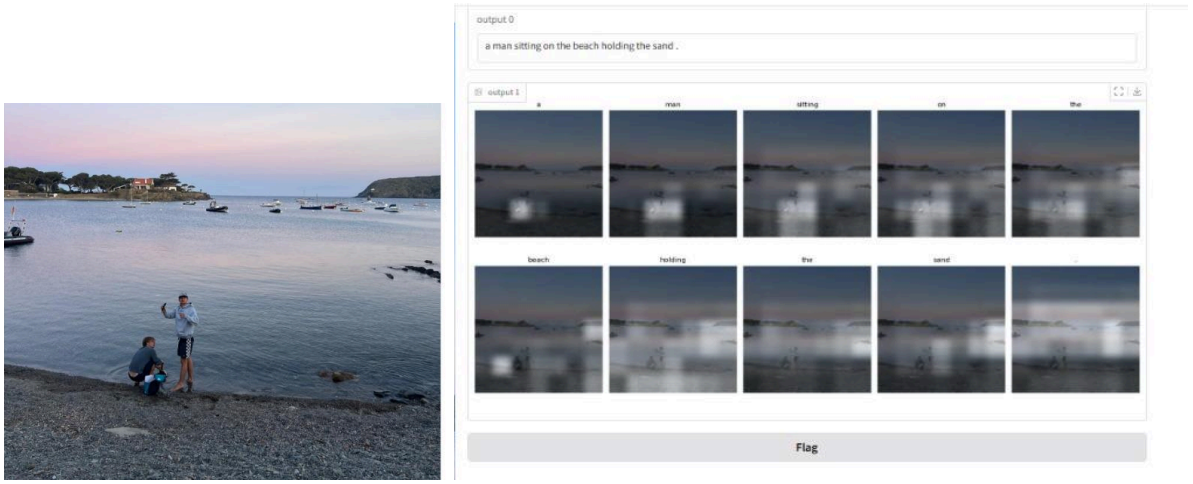

Figure 11: Caption and attention maps of an image of our own generated with model trained with all samples and resnet-152

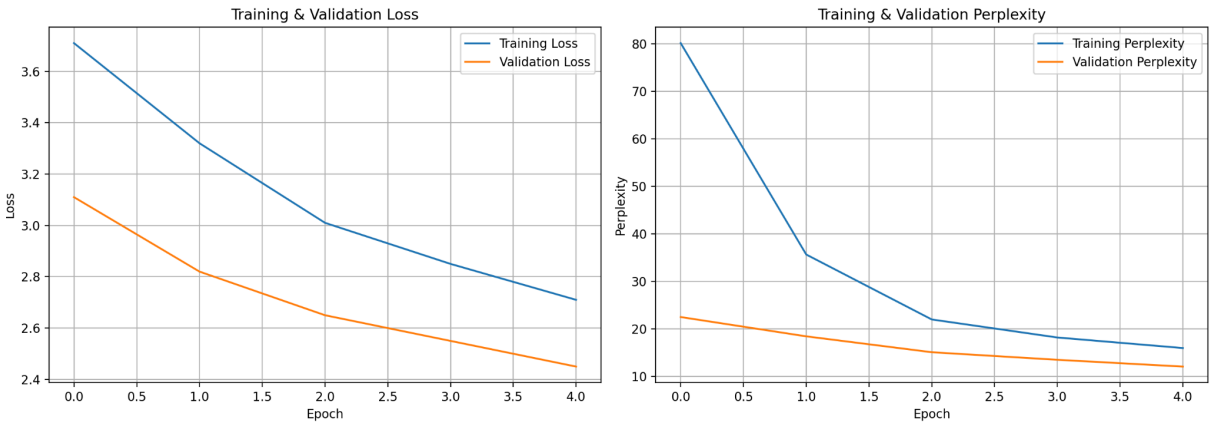

Figure 12: Train and validation loss and perplexity for training with all samples and resnet-50

7. DISCUSSION AND FUTURE WORK

The results of our experiments demonstrate both the potential and limitations of attention-based image captioning using encoder-decoder architectures. Our methodological changes — namely, integrating attention and switching to top-p sampling — were based on the hypothesis that enhancing spatial focus and introducing output variability would lead to more coherent and diverse captions. These expectations were largely met, especially once we scaled up the dataset and increased model capacity.

Figures 3 to 11 illustrate the model's evolution. At early stages — with subsets of 1.5k to 8k images — results were poor: captions were highly repetitive and often generic, and attention maps showed erratic or inconsistent focus. This likely stems from insufficient data diversity and class imbalance, limiting the model's ability to form strong semantic associations. Even with 25k samples, captions improved structurally but remained visually inaccurate, highlighting the limitations of partial training sets.

A turning point came when training on the full MS COCO dataset with ResNet-50 (Figure 7), where fluency and grounding improved significantly. Validation perplexity dropped below 15, and captions better reflected scene context. Yet, syntactic issues persisted, showing that more data alone doesn't guarantee optimal results without richer features.

Replacing ResNet-50 with ResNet-152 further improved semantic grounding. As shown in Figures 9 to 11, this deeper encoder enabled finer visual detail capture and better alignment between vision and language. For instance, while the model in Figure 10 misclassified animals, the attention map correctly highlighted the "grass," indicating functional spatial focus despite labeling errors. Likewise, Figure 11 — on an unseen personal image — produced a plausible caption, showing promising generalization despite occasional quirks.

Nonetheless, several issues persist. The model still shows a tendency to generate high-frequency phrases, a reflection of the dataset's class imbalance (as seen in Figure 2), and sometimes confuses semantically similar concepts. These limitations point to deeper challenges in compositional reasoning and long-tail representation in captioning systems.

Looking forward, several improvements could strengthen both performance and interpretability. First, expanding the evaluation metrics beyond loss and perplexity to include

BLEU or CIDEr would allow for more comprehensive and task-aligned assessment. Second, hyperparameter tuning (e.g., learning rate, dropout, attention dimensionality) was not explored due to computational constraints but could reveal better optimization dynamics. Third, moving from an LSTM decoder to a Transformer-based architecture would enable more flexible modeling of both visual and linguistic dependencies, and has shown strong results in related captioning tasks.

Lastly, while transfer learning from ImageNet proved valuable, incorporating multimodal pretraining or leveraging large-scale vision-language models (e.g., CLIP, BLIP) could offer stronger semantic priors and improve sample efficiency. These directions, however, demand substantial computational resources beyond the scope of our project. Nevertheless, our results show that with careful architectural choices — particularly the use of attention, sampling, and deep encoders — it is possible to build competitive and interpretable captioning systems even under modest resource constraints

8. CONCLUSIONS

- **Attention mechanisms enable spatially-aware captioning:** Integrating attention into the decoder significantly improved the model's ability to focus on meaningful image regions when generating captions. This led to more coherent, contextually grounded descriptions and enhanced interpretability through attention maps.

- **Top-p sampling introduces output diversity:** Switching from deterministic argmax decoding to top-p sampling allowed the model to produce varied captions for the same image. This better reflects human-like language behavior and avoids rigid, repetitive outputs, improving the naturalness of the results.

- **Caption quality highly depends on dataset scale and encoder depth:** Training on small subsets resulted in repetitive, generic captions with poor attention focus. Notable improvements only emerged when using the full dataset and upgrading the encoder, underscoring the role of data scale and visual richness.

- **Deeper encoders improve semantic grounding:** The ResNet-152 backbone captured finer visual details, yielding better object recognition and context-aware language without changing the decoder architecture.

- **The model generalizes to unseen images:** When tested on external, non-COCO images, the model produced structurally sound and generally accurate captions. While not flawless, this shows promising robustness beyond the training distribution.

- **Dataset biases still influence predictions:** Frequent phrases and dominant object types appear often, showing the model's reliance on dataset frequency over true visual grounding. This reflects persistent long-tail and bias challenges.

- **Results remain imperfect, with room for improvement:** The model occasionally produces incorrect, vague, or mismatched captions, especially for complex or rare scenes. These errors point to the need for better semantic understanding and broader training diversity.

9. REFERENCES

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. Proceedings of the 32nd International Conference on Machine Learning (ICML). https://arxiv.org/abs/1502.03044

2. Sharma, A., Pavan, P., & Narayanan, R. (2024). *A Survey on Image Captioning Techniques and Recent Advances*. arXiv preprint arXiv:2404.18062. https://arxiv.org/abs/2404.18062

3. Wang, W., Xie, E., Li, X., Ma, X., Lu, T., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., & Ding, E. (2022). *SimVL: Simple Visual-Linguistic Pretraining with Vision Transformers*. arXiv preprint arXiv:2203.15350. https://arxiv.org/abs/2203.15350

4. Yunjey. (2018). *Image Captioning (PyTorch tutorial repository)*. GitHub. https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning

5. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1908.02265

6. Li, J., Li, D., Xiong, C., & Hoi, S. C. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. Proceedings of the 39th International Conference on Machine Learning (ICML). https://arxiv.org/abs/2201.12086