Impact of Dataset Size and Distillation Techniques on Image Captioning Performance: An Empirical Study

Srushti Sangawar and Arunava Ghosh

Course CSCI 5922, University of Colorado Boulder

Abstract. Image captioning is a task that combines computer vision and natural language processing, in which a model generates descriptive text from an input image. High-performance captioning systems have traditionally depend on computationally intensive models and big datasets such as MSCOCO. But building such models from the ground up requires a lot of resources, and people without a lot of processing power frequently cannot afford it. This study aims to investigate how well pre-trained architectures, like ResNet-50, GIT, and GPT-2, work in combination with dataset distillation strategies to minimize the large volumes of training data without substantially affecting model performance. We use random selection and gradient-based distillation as sampling techniques to assess models trained on distilled subsets of different sizes (25%, 50%, 75%, and 100%), and we evaluate their performance using BLEU and CIDEr measures.

Keywords: Image Captioning, Dataset Distillation, Gradient-Based Selection, Random Sampling, ResNet-50 + LSTM, CLIP + GPT-2, GIT Model, MSCOCO Dataset, BLEU Score, CIDEr Score, Deep Learning, Vision-Language Models, Pretrained Transformers, Data Efficiency, Model Evaluation, Caption Quality, Low-Resource Training

1 Introduction

Image captioning [1] is about teaching computers to look at an image and describe it in words, similar to how a person might say "a cat sitting on a couch" when shown a photo. It blends computer vision and natural language processing and has been studied through models like Show and Tell or Show, Attend and Tell, which combine convolutional neural networks (CNNs) with recurrent neural networks (RNNs). In recent years, transformer-based models like GIT have taken this further by generating more fluent and context-aware captions. However, these models are often very large, require huge datasets like MSCOCO, and take a long time to train, which makes them hard to use in resource-limited situations. That's where dataset distillation comes in. Instead of training on all available data, distillation tries to find the most informative samples so that models can learn faster and with less data. This is especially helpful in real-world environments like mobile apps, embedded devices, or schools and clinics, where computing resources are limited. Smaller models like GPT-2 can be used for captioning by converting image features into a form the language model understands. This makes it possible to build lightweight captioning systems that are still quite accurate and useful. Our project focuses on making these models more efficient by combining pre-trained architectures with data distillation strategies. The end goal is to create accessible captioning tools that could help with things like better image search, personalized recommendations, or even assistive tech for visually impaired users.

Even with all the recent progress, there are still some roadblocks to using image captioning widely. The best models need a lot of labeled data and powerful hardware to train properly, which isn't always available. Models like GIT perform really well but are hard to run in settings where speed or cost matters. Past research has mostly explored distillation for simpler tasks like classification, where the goal is to pick a label from a list. For image captioning, which involves generating full sentences with proper grammar and meaning, distillation is more challenging and less studied.

In our work, we explore how different captioning models respond to reduced training data using two methods: random sampling and gradient-based selection. We test these across a range of dataset sizes to see how performance changes, using BLEU and CIDEr scores to measure quality, along with training and inference time. Our goal is to understand how much data is really needed, how distillation can help, and which models work best when resources are limited.

2 Related Work

Dataset Distillation for Efficient Training: Our work is different from these prior efforts in several key aspects. Yu et al. (2023) [2] provide a comprehensive survey of dataset distillation techniques, classifying them based on optimization strategies, data modality, and application domains. However, their work remains largely theoretical and review-based, without conducting new experimental comparisons across distillation methods. In contrast, our approach focuses on an empirical evaluation of gradient-based and random selection methods to construct distilled datasets of varying sizes (25%, 50%, 75%, and 100%). This allows us to systematically analyze the trade-offs between dataset size, selection strategy, and downstream model performance. Furthermore, Lopes et al. [3] propose a data-free knowledge distillation method that synthesizes inputs using prior information from pre-trained models. Their method is particularly useful in privacy-sensitive scenarios where training data is unavailable. However, our work is rooted in the practical context where full datasets are available, and the goal is to reduce training data size for efficiency, without introducing synthetic inputs or relying on student-teacher architectures. By comparing distilled subsets obtained through different strategies, our work offers insights for practitioners aiming to balance training cost with accuracy.

Image Captioning with Pre-trained Models: The paper "Show and Tell: A Neural Image Caption Generator" [4] merges convolutional neural networks for image feature extraction and LSTM networks to produce natural language captions. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" [5] provides information on how enabling the model to dynamically focus on the right areas of the image while generating each word in the caption. The research shown in the paper "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" [6] improves caption quality using bottom-up and top-down attention, where object-level features learned using Faster R-CNN are selectively focused on while decoding.

In comparison to these existing works, our work is looking to improve efficiency and scalability instead of focusing on architectural complexity. Instead of leveraging the attention-heavy models or relying on the object detectors themselves, our work uses pretrained models like ResNet-50 to effectively extract features from images. Instead of attempting to enhance the quality of captions depending on the complex mechanisms of attention, our work promotes training optimization, simplifying the dataset size with distillation and adding examples to it. Because it reduces training time and computational demand at the expense of competitive performance, this is especially well-suited to environments with limited resources. In image captioning research, there has been a noticeable shift from model-centric to data-centric innovation with the use of pre-trained models in conjunction with data-centric optimization strategies such as dataset distillation.

3 Methodology

To study the effect of reduced training data, we used a 5,000-image subset from the MSCOCO 2017, each image annotated with five human-written captions. All images were resized to 224×224 pixels, and captions were tokenized, numericalized, and padded to a maximum length of 32 tokens. As a baseline, we are using random sampling, where we constructed training subsets by randomly selecting 25%, 50%, 75%, and 100% of the full dataset using Python's random.sample() over the list of image IDs. This approach simulates low-resource environments and provides a naïve benchmark to compare against more targeted distillation.

To evaluate the effects of dataset distillation, we trained and compared three distinct image captioning models. The first architecture combined a pretrained ResNet-50 encoder with an LSTM decoder. Specifically, we removed the classification head of ResNet-50 and used the extracted 2048-dimensional feature vector, which was projected down to 256 dimensions using a linear layer. This vector was then passed as the initial input to a single-layer LSTM with a hidden size of 256. The LSTM output was mapped to vocabulary scores through a fully connected layer, and the model was trained using the Adam optimizer with a learning rate of 1e-4 for 20 epochs using sparse categorical cross entropy as loss, where padding tokens were ignored in the loss calculation.

The second model in our experimental pipeline combined a CLIP-based visual encoder with a GPT-2 language decoder. We utilized the CLIP ViT-B/32 architecture to extract 512-dimensional embeddings for each image in the

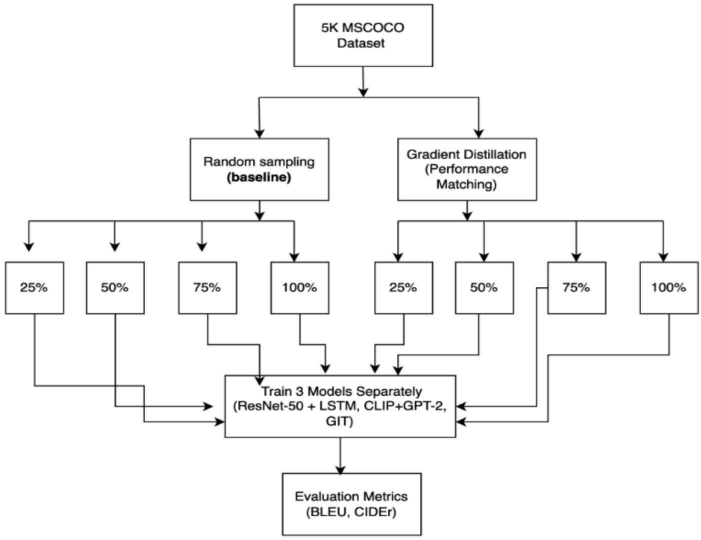

Fig. 1. A flowchart showing an overview of the experiments

dataset. To enable compatibility with the GPT-2 decoder, which operates in a 768-dimensional embedding space, we employed a learnable linear projection layer that mapped the CLIP image features to the GPT-2 token embedding space. During training, the projected image embedding was prepended as the first token in the caption sequence, effectively grounding the textual generation in visual context. The tokenized caption was then passed to a pretrained GPT2LMHeadModel from HuggingFace's transformers library. We fine-tuned both the projection layer and GPT-2 weights end-to-end using the AdamW optimizer with a learning rate of 5e-5 for 20 epochs. This architecture allowed us to repurpose a text generation model for image captioning tasks with minimal structural changes, enabling efficient multimodal learning.

Finally, we used the GIT model, which integrates both a Vision Transformer (ViT) encoder and a transformer-based language decoder into a unified, end-to-end architecture for image captioning. Unlike the other models, GIT does not require separate feature extraction or projection layers; instead, it takes raw image tensors as input and internally handles multimodal alignment. Preprocessing was handled using the GitProcessor, which tokenizes captions and applies the appropriate image transformations to align with GIT's pretrained pipeline. Due to the model's scale and computational demands, we fine-tuned GIT for 20 epochs using the AdamW optimizer with a learning rate of 5e-5. Evaluation was conducted using BLEU-1 through BLEU-4 and CIDEr scores, consistent with our methodology across all experiments. This architecture required minimal intervention and demonstrated high performance, making it well-suited for captioning tasks.

To explore performance with smarter data reduction, we implemented gradient-based distillation using the same subset of MSCOCO dataset. In this approach,

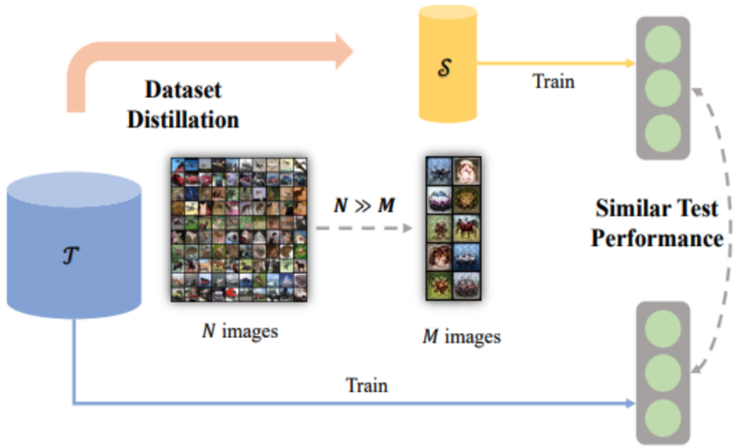

Fig. 2. An overview showing the dataset distillation process

we first trained the model on a small subset and computed per-sample gradient norms and storing this information across batches. We then ranked all samples based on these gradient norms and selected the top 25%, 50%, and 75% most informative examples to create distilled datasets. This ensures we retain examples that cause the model to learn more, akin to keeping only the hardest or most educational questions in a study guide. Each distilled subset was used to train the same three models described above. All models were trained using the same architecture, batch sizes (typically 64), and learning rates as in the random baseline to ensure fair comparison. The BLEU-1 to BLEU-4 and CIDEr scores were used to evaluate caption quality, while training time and inference efficiency were monitored to assess computational impact. This dual comparison allowed us to assess not just model performance but also whether smarter data curation could save compute while preserving or improving quality.

Each distilled subset was used to train the same three models described above. All models were trained using the same dataset sizes, batch sizes (typically 64), and learning rates as in the random baseline to ensure fair comparison. The BLEU-1 to BLEU-4 and CIDEr scores were used to evaluate caption quality, while training time and inference efficiency were monitored to assess computational impact. This dual comparison allowed us to assess not just model performance but also whether smarter data curation could save compute while preserving or improving quality. All code will be made publicly available to ensure reproducibility.

Table 1. Comparison of Image Captioning Models

Feature	ResNet-50 + LSTM	GIT Model	CLIP + GPT-2
Image Encoder	ResNet-50 (pre-trained CNN)	Transformer-based Vision Encoder (ViT-like, BERT-style pretraining)	CLIP pretrained on image-text pairs
Text Decoder	LSTM RNN	Multimodal Transformer Decoder	GPT-2 Transformer Decoder
Image-Text Connection	Image features fed as initial LSTM input	Unified vision-language model	Linear projection maps image features to GPT-2 space
Training Type	Train LSTM on top of fixed ResNet-50 features	Fine-tune full model (encoder + decoder)	Fine-tune GPT-2 with learnable image token
Pretraining Used	ResNet-50 ImageNet pretrained	Pretrained on large image-text datasets	CLIP and GPT-2 both pretrained
Strength	Simple and fast; Good with small data	Strong multimodal understanding	Strong language modeling capabilities
Weakness	LSTM less powerful than Transformers	Heavy model; more compute needed	Needs strong image embedding quality

4 Experiments

In this project, we propose the following experiments to test how effective of dataset distillation is in improving image captioning models. Also, we will be looking into how different models act under the experiment setup for image captioning.

Experiment 1: Comparing Distillation Methods on Full Dataset Using ResNet-50

– **Main purpose:** In this experiment, we trained a ResNet-50 [7] + LSTM captioning model on the subset of MSCOCO dataset (5,000 images) to compare two distillation methods: random sampling and gradient-based selection. The random approach served as a baseline for our experiment. Gradient-based selection prioritized images with the highest per-sample gradient norms, assuming that these examples would be more informative. The model was trained over 20 epochs using the Adam optimizer (learning rate = 1e-4), with BLEU-1 to BLEU-4 and CIDEr scores used for evaluation.
– **Evaluation Metric(s):**
  - **BLEU:** To measure the n-gram precision of the generated captions against reference captions, providing a standard metric for caption quality [8] [9].

  - **CIDEr:** To evaluate the consensus between generated and reference captions, which is particularly useful for human-like captioning tasks [10] [11].
  - **Training Time:** To assess computational efficiency and determine how much faster training can be completed using distilled datasets.
- **Results:**
Results showed that gradient-based distillation slightly outperformed random sampling across all metrics. For instance, BLEU-4 improved from 0.098 (random) to 0.105 (gradient), and CIDEr rose from 0.246 to 0.259. Although training speed was slightly slower for gradient-based selection (27.46 vs. 28.37 iters/sec), the generated captions were more detailed and semantically accurate. This suggests that examples taken with higher focus during training can improve generalization.
We also observed a consistent trend where gradient-based selection led to better performance than random sampling, indicating that not all examples contribute equally to learning. By focusing on the important samples, the model was able to extract richer information, leading to higher quality captions despite identical data volumes.
However, this experiment leaves open some questions about scalability and broader generalization. Would the same benefit hold for larger datasets or with different architectures? Also, gradient-based selection is computationally expensive, requiring extra forward and backward passes per sample, so future work should explore whether similar gains can be achieved using cheaper approximations (e.g., entropy-based filtering or active learning). A promising next step is to analyze the interaction between model complexity and distillation methods.

Model	**Distillation Method**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**
ResNet50	Random	0.427	0.273	0.166	0.098	0.246
ResNet50	Gradient	0.455	0.292	0.177	0.105	0.259

Table 2. Comparison of random and gradient-based distillation methods using ResNet50.

Experiment 2: Varying the Size of Distilled Datasets Using ResNet-50

- **Main purpose:** Building on the first experiment, this setup explores how reducing the size of the training dataset affects model performance. We used the same ResNet-50 + LSTM architecture and tested it across four dataset sizes, 25%, 50%, 75%, and 100% of the MSCOCO subset (5,000 images). For each size, we compared both random sampling and gradient-based distillation while holding all hyperparameters constant: 20 training epochs, Adam optimizer (learning rate = 1e-4), and sparse categorical cross-entropy loss. The goal was to understand how much data is truly necessary for acceptable captioning performance, especially under limited computational resources.

– **Evaluation Metric(s):**
  - **BLEU:** To track changes in caption precision as dataset size decreases [9].
  - **CIDEr:** To assess how closely captions generated from smaller datasets match those from full datasets [11].
  - **Training Time:** To evaluate computational cost reduction as the dataset size is reduced.

– **Results:**
The results tells us that captioning performance improves with more data, but most of the gains become stagnant after 75%. For example, BLEU-4 for gradient-based distillation rises from 0.099 at 25% to 0.109 at 75%, but barely improves beyond that. Interestingly, the 75% gradient-based model nearly matches the 100% performance, with CIDEr increasing from 0.251 to 0.292. This suggests that smaller, carefully selected datasets can approach full-dataset performance while reducing training time and compute costs. This could be vital for training in environments where resources are limited. While the improvements from increasing the dataset are consistent, the marginal gains diminish past the 75% mark. This opens up an important direction for future work: Can we combine advanced distillation techniques with model pruning or quantization to further optimize performance? Additionally, these results raise questions about the optimal balance between dataset size and model complexity, especially when larger models are involved.

Model	Distillation Method	Dataset Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr
ResNet50	Random	25%	0.433	0.277	0.167	0.096	0.251
ResNet50	Random	50%	0.428	0.273	0.165	0.097	0.247
ResNet50	Random	75%	0.426	0.272	0.165	0.097	0.245
ResNet50	Random	100%	0.427	0.273	0.166	0.098	0.246
ResNet50	Gradient	25%	0.459	0.299	0.183	0.099	0.251
ResNet50	Gradient	50%	0.446	0.283	0.172	0.101	0.258
ResNet50	Gradient	75%	0.446	0.282	0.170	0.109	0.292
ResNet50	Gradient	100%	0.455	0.292	0.177	0.105	0.259

Table 3. Performance comparison across different dataset sizes and distillation methods using ResNet-50.

Experiment 3: Comparing Different Pretrained Models

– **Main purpose:** In this final experiment, we evaluated the impact of model architecture on image captioning performance by training three different models under identical experimental conditions. We kept the training setup, dataset, and distillation techniques constant using 75% of the MSCOCO subset based on insights from Experiment 2 and varied only the model architecture. Specifically, we compared ResNet-50 + LSTM (used as our baseline), CLIP [12] + GPT-2 [13] (where image embeddings from CLIP were mapped into GPT-2's language space using a learnable projection), and GIT [14] (a powerful, transformer-based vision-language model pretrained end-to-end for captioning tasks).

– **Evaluation Metric(s):**
Like in the previous experiments we will be evaluating our experiment via the BLEU, CIDEr scores as well as training time. In addition to this we will evaluate the generated caption manually to understand how accurately the model is able to explain the image through the generated captions.

– **Results:**
The results clearly demonstrated that model architecture had the strongest influence on captioning performance. GIT consistently achieved the highest scores, with a BLEU-4 of 0.788 and CIDEr of 2.133 under gradient-based distillation, followed by GPT-2 + CLIP (BLEU-4: 0.624, CIDEr: 1.792). ResNet-50 + LSTM, while much faster to train (23.34 iterations/sec), showed the weakest captioning ability (BLEU-4: 0.097–0.105, CIDEr: 0.25). Notably, gradient-based distillation provided slight improvements across all models, but the gap between architectures was far more impactful than the gap between distillation methods. A general trend we observed was that transformer-based models like GIT and GPT-2 benefited more from distilled data than the ResNet-based baseline, possibly because they are better at leveraging contextual signals when trained on more informative samples. The GPT-based model, while not as powerful as GIT, showed strong results with significantly lower training time than GIT, suggesting a useful balance for mid-resource scenarios. GIT's superior performance aligns with its design, it's pretrained end-to-end for vision-language tasks, allowing it to generalize better, even on smaller or distilled datasets. However, some important questions remain open. For instance, while we manually inspected caption outputs, we did not formally evaluate fluency or factual correctness beyond BLEU/CIDEr metrics, which may not always align with human judgment. Additionally, our experiments were conducted on a relatively small subset (5K images), and it is unclear how these trends would scale on the full MSCOCO or other datasets. Future directions include testing on larger datasets, applying other distillation techniques, and exploring multi-modal fine-tuning of GIT to reduce its training time without sacrificing quality.

Fig. 3 illustrates how BLEU-1 scores vary with dataset size across different models and distillation methods. It shows that GIT achieves the highest scores overall, and that gradient-based distillation slightly boosts performance, especially for smaller models, with most gains plateauing after 75% dataset size.

Fig. 4 compares CIDEr scores across models and distillation techniques as dataset size increases. GIT consistently outperforms other models, while gradient-based distillation provides a modest improvement for GPT and ResNet-50. The

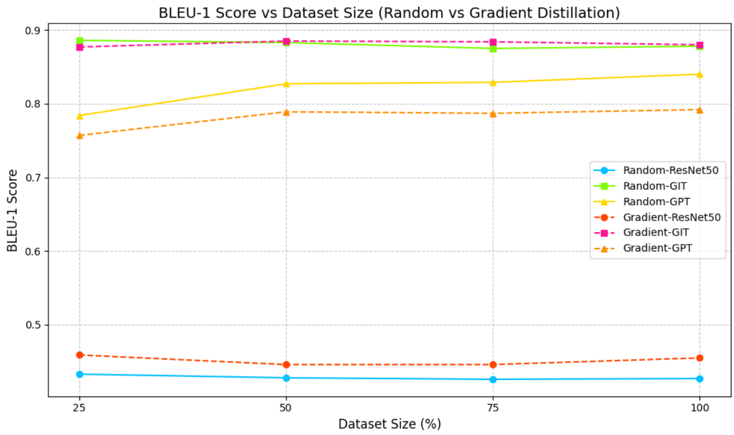

Fig. 3. BLEU-1 Score vs Dataset Size across three different models

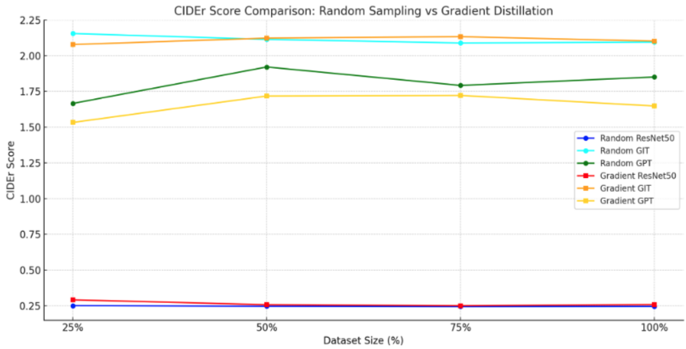

Fig. 4. A flowchart showing an overview of the experiments

Model	Distillation Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	Time (iters/sec)
ResNet50	Random	0.426	0.272	0.165	0.097	0.245	23.34
GIT	Random	0.875	0.828	0.796	0.773	2.089	1.45
GPT	Random	0.829	0.745	0.676	0.624	1.792	10.46
ResNet50	Gradient	0.446	0.282	0.170	0.099	0.251	23.34
GIT	Gradient	0.884	0.841	0.811	0.788	2.133	1.46
GPT	Gradient	0.787	0.688	0.610	0.548	1.722	10.46

Table 4. Comparison of BLEU, CIDEr scores, and training speed across models and distillation methods.

performance largely saturates beyond 75% dataset size, reaffirming that high-quality subsets can approximate full-data performance.

Fig. 5. A sample test image

The test image fig.5 depicts a dining room scene with a table, a visible fireplace, and a person standing in the area. An ideal caption would reference these key components, for example, “a woman standing near a table in front of a fireplace.” While the GIT model generates the most contextually relevant caption, it still omits some critical details. The ResNet-50 model repeats objects redundantly, and CLIP + GPT-2 produces a vague and incomplete output. This discrepancy highlights a key limitation of automatic evaluation metrics. As a result, models may perform well quantitatively but still struggle to generate comprehensive and human-like captions in real-world scenarios.

Model	Generated Caption
ResNet-50 + LSTM	image of a fireplace in a table with a table and a fireplace
GIT	a woman stands in the dining area at the table
CLIP + GPT-2	A window and home with yellow in the United States of America, a man was walking

Table 5. Captions generated by different models for the same input image as shown in Fig 5.

5 Conclusions

This study tells us how effective the technique of dataset distillation is in enhancing the efficiency of image captioning models under constrained data and computational resources. We conducted systematic evaluation of three pretrained architectures: ResNet-50 with LSTM, CLIP with GPT-2, and GIT. These models were tested across two strategies: random sampling and gradient-based selection. Our findings show that gradient-based distillation provides consistent but modest improvements over random sampling across all dataset sizes and model architectures. Furthermore, performance gains become stagnant beyond the 75 percent data. This indicates that carefully selected subsets can offer performance comparable to full-scale training while substantially reducing computational overhead.

Among the evaluated models, GIT achieved the highest BLEU and CIDEr scores, showcasing its strength in capturing visual-semantic alignment. The CLIP and GPT-2 combination provided a good trade-off between performance and efficiency, while ResNet-50 with LSTM established a solid baseline. However, manual inspection of generated captions revealed a limitation in standard evaluation metrics. Despite high BLEU and CIDEr scores, some captions lacked key semantic details. This highlights the inadequacy of relying solely on n-gram-based metrics, which may overestimate caption quality in real-world scenarios.

From the point of view of ethics for this work, we have used open-source data and pretrained models for our research. The thing that could negatively impact this could be biased training of model which could result in the model providing improper captions. This could have a negative impact on the application as well as the user. If we can maintain an unbiased mode of training and regularly monitor the data that is fed to the model for training then this research and its further studies could really have a positive impact.

Future directions for this work could include exploring more robust evaluation frameworks that combine automatic and human-centered assessments to better capture caption relevance. In addition to this, integrating dataset distillation with model compression methods such as pruning, quantization, or knowledge distillation could further enhance the deployability of captioning systems on low-power devices. Another avenue for further expanding on our work is extending this methodology to multilingual and low-resource settings, or using synthetic data augmentation to increase diversity without requiring additional labeling.

Finally, applying these techniques to other vision-language tasks, including visual question answering and image-text retrieval, may provide broader insights into the benefits and limitations of distillation strategies.

Overall, our research supports the development of scalable, efficient, and accessible image captioning systems. It demonstrates that smart data selection combined with strong architectures can result in high-quality outputs even in low-resource environments.

References

1. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. arXiv preprint arXiv:1906.05963 (2020) [Online]. Available: https://arxiv.org/abs/1906.05963.
2. Ruonan Yu, Songhua Liu, X.W.: Dataset distillation: A comprehensive review (2023)
3. Raphael Gontijo Lopes, T.S.: Data-free knowledge distillation for deep neural networks (2017)
4. S. Bengio, D.E.: Show and tell: A neural image caption generator (2015)
5. Ruslan Salakhutdinov, Richard Zemel, Y.B.: Show, attend and tell: Neural image caption generation with visual attention (2015)
6. Peter Anderson, L.Z.: Bottom-up and top-down attention for image captioning and visual question answering (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
8. Singh, S.: Evaluation and fine-tuning for image captioning models - a case study (2024) https://www.labellerr.com/blog/image-captioning-evaluation-and-fine-tuning.
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), pp. 311–318 (2002) [Online]. Available: https://doi.org/10.3115/1073083.1073135.
10. Ramakrishna Vedantam, C. Lawrence Zitnick, D.P.: Cider: Consensus-based image description evaluation (2015)
11. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. arXiv preprint arXiv:1411.5726 (2015) [Online]. Available: https://arxiv.org/abs/1411.5726.
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) [Online]. Available: https://arxiv.org/abs/2103.00020.
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Technical Report (2019) [Online]. Available: https://cdn.openai.com/better-language-models/language$_{m}odels_{a}re_{u}nsupervised_{m}ultitask_{l}earners.pdf$.
14. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022) [Online]. Available: https://arxiv.org/abs/2205.14100.