Image Caption Generation using a Vision Transformer and a Transformer Decoder

Hamza Ali Khan
School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK
Student ID: 001345840 Email: hk2108k@gre.ac.uk

Abstract. This report presents an image captioning system with a frozen pretrained Vision Transformer (ViT) encoder and a Transformer decoder trained from scratch on Flickr8k. The pipeline covers exploratory data analysis, text and image preprocessing, image-level train/validation/test splitting, a deliberately simple baseline decoder, and an Optuna-tuned model (100 trials). Caption quality is evaluated with BLEU, METEOR, and CIDEr under greedy and beam-search decoding. Quantitative results are complemented by qualitative examples, a manual error taxonomy, and ViT-based attention heatmaps. With beam search the final system reaches BLEU-4 = 0.1852, METEOR = 0.1941, and CIDEr = 0.4635, clearly outperforming the baseline greedy decoder while remaining feasible on a single GPU.
Keywords: Image Captioning, Vision Transformer, Transformer Decoder, Beam Search, Hyperparameter Optimisation, Flickr8k

1 Introduction

Image captioning automatically generates natural-language descriptions of visual scenes from paired image–text data. Modern systems follow an encoder–decoder paradigm in which a vision backbone extracts image features and a language model decodes them; early CNN+LSTM models have largely been replaced by Transformer architectures for both vision and language [13, 4].

This work studies a compact ViT+Transformer captioner in a coursework setting. A pretrained ViT-B/16 encoder is frozen to provide stable visual features, and a Transformer decoder is trained from scratch on the Flickr8k dataset. The focus is on decoder design, hyperparameters, and decoding strategy rather than on training a large vision backbone, which would be unrealistic for 8 091 images. Key design choices—freezing the encoder, discarding accuracy as a metric, and using one caption per image—are motivated by the problem domain and resource constraints.

The study addresses three questions: (i) how well a simple ViT+Transformer baseline performs when only the decoder is trained; (ii) to what extent automatic hyperparameter optimisation with Optuna can improve this decoder; and (iii) whether beam search with a small beam width yields better BLEU, METEOR and CIDEr scores than greedy decoding.

2 Background and Problem Formulation

Transformers are now the dominant architecture for sequence modelling in language and vision [13]. Self-attention lets each token attend to all others in parallel, improving long-range dependency modelling over RNNs. Vision Transformers (ViTs) extend this to images by splitting them into fixed-size patches and treating each patch as a token [4]. When pretrained on large datasets such as ImageNet, ViTs learn general-purpose visual features that transfer well to smaller downstream tasks.

Image captioning can be formalised as learning $P(y_{1:T} \mid I)$, the probability of a token sequence $y_{1:T}$ given an image I. In this project a frozen pretrained ViT maps I to a feature vector, and a Transformer decoder models the conditional distribution over tokens. Because multiple human captions may be equally valid for the same image, sequence-based metrics such as BLEU, METEOR, and CIDEr are used instead of accuracy or confusion matrices [9, 3, 14].

The system starts from a compact baseline decoder and then improves it via hyperparameter tuning and beam search. The aim is not state-of-the-art performance (which would require web-scale pretraining [8]) but a reproducible, well-analysed model under realistic compute constraints.

3 Methods and Exploratory Data Analysis

3.1 Dataset and Exploratory Data Analysis

The experiments in this report are based on the publicly available Flickr8k dataset, introduced by Hodosh et al. [5] and accessed via a Kaggle mirror of the original release [6]. Flickr8k consists of 8 091 images, each annotated with five human captions, for a total of 40 455 descriptions. An initial EDA step verified dataset integrity and guided subsequent design choices.

Duplicate and missing entries were checked: fewer than 1% of rows were exact duplicates and were removed, and all images referenced in the caption file were present on disk. Caption lengths ranged from 2 to 40 words, with a mean of approximately 12 words and the majority (about 97%) shorter than 20 tokens. This motivated setting the maximum caption length to 22 tokens, leaving a small margin while avoiding unnecessary padding. A word-frequency analysis confirmed that the dataset is biased towards human- and animal-centric scenes; high-frequency content words include *man*, *woman*, *dog*, and *people*. This bias is noted later as a limitation for generalisation.

All images are resized to 224×224 pixels for ViT preprocessing.

3.2 Preprocessing and Data Splits

Text preprocessing aimed to reduce noise while preserving semantics. All captions were lowercased, punctuation and non-alphanumeric characters were removed, and multiple spaces were

collapsed. Each caption was then tokenised at word level and wrapped with special tokens <start> and <end>.

To prevent data leakage, splits were made at the *image level*: all five captions for a given image reside in a single split only. The dataset was partitioned into 80% training, 10% validation, and 10% test images. A frequency-based vocabulary was constructed from the training captions only, keeping the 8 000 most frequent words plus four reserved tokens (<pad>, <start>, <end>, <unk>). Out-of-vocabulary rates on validation and test captions were around 1%, indicating that the vocabulary size is sufficient. Captions were converted to integer sequences and padded or truncated to a fixed length, enabling batched training.

During training, a single caption per image was used. This simplifies the dataloader and ensures each training step sees a unique pairing of image and caption. The drawback is that the decoder is exposed to less linguistic variety than the dataset potentially offers; this is highlighted as a limitation and a clear direction for future work.

4 Model Architecture and Optimisation

4.1 Frozen ViT Encoder

The image encoder is a pretrained ViT-B/16 model [4] loaded from a public model hub. It divides each 224×224 image into 16×16 patches, produces 768-dimensional patch embeddings, and processes them with multi-head self-attention. In this work the encoder is *frozen* (`requires_grad = False`). This choice is motivated by the small size of Flickr8k: Vision Transformers typically require large-scale data to fine-tune reliably, and partial unfreezing on small datasets is known to cause rapid overfitting and optimisation instability even when only the last blocks are trained [7, 12]. Freezing the encoder therefore preserves robust ImageNet-pretrained features while keeping the experiment focused on decoder design, hyperparameters, and decoding strategy.

4.2 Transformer Decoder

The caption decoder is a stack of Transformer blocks [13] with model dimension d_{model} and feed-forward dimension d_{ff}. Each block consists of:

- masked self-attention over previously generated tokens;
- cross-attention over the projected image embedding;
- a position-wise feed-forward network with ReLU activation;
- residual connections and layer normalisation.

An embedding layer maps token IDs to d_{model}-dimensional vectors, and sinusoidal positional encodings inject information about token order. The final linear layer followed by softmax produces a distribution over vocabulary items at each decoding step.

4.3 Training and Hyperparameter Optimisation

Training uses the AdamW optimiser with an initial learning rate of 1×10^{-4}, mini-batches of size 64, and cross-entropy loss over target tokens, ignoring <pad> positions:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, I), \tag{1}$$

where y_t is the ground-truth word at time step t and I denotes the image embedding. The baseline decoder is trained for 10 epochs; the optimised decoder for up to 20 epochs with early stopping based on validation loss (patience $= 4$). Gradient clipping prevents exploding gradients.

The baseline captioning model is deliberately simple, with $d_{\text{model}} = 256$, $n_{\text{heads}} = 1$, N=1 decoder layer, $d_{\text{ff}} = 1024$, dropout 0.1, and learning rate 1×10^{-4}. Here d_{model} is the decoder hidden size, n_{heads} the number of attention heads, N the number of decoder layers, d_{ff} the inner feed-forward width, *dropout* the regularisation rate, and *lr* the learning rate. This minimalist configuration provides a coherent reference model rather than an aggressively tuned system.

To obtain a stronger model, Optuna [1] is used for hyperparameter optimisation. Optuna performs Bayesian optimisation with pruning: each trial trains the decoder for a small fixed number of epochs, reports validation loss, and poorly performing trials are terminated early. The search space includes

$$\begin{aligned} d_{\text{model}} &\in \{256, 320, 384\}, & n_{\text{heads}} &\in \{2, 4\}, \\ N &\in \{1, 2, 3\}, & d_{\text{ff}} &\in \{1024, 1536, 2048\}, \\ \text{dropout} &\in [0.05, 0.25], & \text{lr} &\in [2 \times 10^{-4}, 6 \times 10^{-4}]. \end{aligned}$$

A budget of 100 trials was chosen as a compromise between exploration and runtime on a T4 GPU. The best trial yields

$$\begin{aligned} d_{\text{model}} &= 320, & n_{\text{heads}} &= 4, & N &= 3, \\ d_{\text{ff}} &= 2048, & \text{dropout} &= 0.0614, & \text{lr} &= 5.368 \times 10^{-4}, \end{aligned}$$

which defines the final "optimised" decoder.

4.4 Decoding and Evaluation Metrics

At test time the model generates captions autoregressively, producing one word at a time conditioned on the previously generated words and the image embedding. Two decoding strategies are used:

- **Greedy search** (k=1): at each step the most probable token is selected. This is fast and simple but myopic and may miss globally better sequences.
- **Beam search** with width k=3: the decoder keeps three partial hypotheses at each step and expands the most likely continuations. A beam of k=3 is a well-established compromise between caption quality and computational cost [15, 11]. Prior work shows that most metric gains occur for $k \in [3, 5]$, with diminishing returns and higher latency beyond this range. For this resource-constrained study, k=3 provided an effective balance.

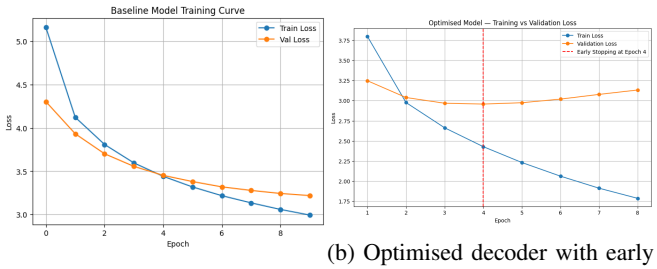
Because captioning admits multiple valid references, accuracy and confusion matrices are not informative. Instead, evaluation uses standard captioning metrics on the test split:

- **BLEU-n** (n=1 . . . 4) measures modified n-gram precision between generated and reference captions with a brevity penalty [9].
- **METEOR** aligns words using stem and synonym matching and combines precision and recall [3].
- **CIDEr** computes a TF–IDF weighted cosine similarity between candidate and reference captions, emphasising informative n-grams [14].

5 Results

5.1 Training Dynamics

Figure 1 shows training and validation losses for the baseline and optimised decoders. The baseline model exhibits a smooth decrease in



(a) Baseline decoder. (b) Optimised decoder with early stopping.

Figure 1: Training vs. validation loss for baseline and tuned decoders.

training loss from 5.16 to 2.99 and validation loss from 4.30 to 3.22 over 10 epochs, with no severe divergence between the curves. This suggests that the model learns a meaningful mapping from images to captions without obvious overfitting.

The optimised model, trained with Optuna-selected hyperparameters, starts at a lower initial validation loss and reaches a minimum of 2.96 around epoch 4 before early stopping triggers. Training loss drops from 3.80 to around 1.76. The consistent improvement in validation loss over the baseline confirms that the tuned configuration is more effective at modelling the caption distribution, although both models remain in a mild underfitting regime.

5.2 Quantitative Evaluation

Table 1: Test-set captioning metrics on Flickr8k.

Metric	Baseline (greedy)	Optimised (greedy)	Optimised (beam)
BLEU-1	0.5458	0.5477	**0.5826**
BLEU-2	0.3723	0.3702	**0.4022**
BLEU-3	0.2477	0.2421	**0.2740**
BLEU-4	0.1671	0.1596	**0.1852**
METEOR	0.1821	0.1917	**0.1941**
CIDEr	0.4165	0.4084	**0.4635**

Table 1 summarises test performance for three systems: the baseline decoder with greedy decoding, the Optuna-tuned decoder with greedy decoding, and the tuned decoder with beam search (k=3). All models are evaluated using BLEU-1–4, METEOR, and CIDEr against the full set of five human reference captions.

The optimised greedy decoder attains a slightly lower validation loss than the baseline (Figure 1), but its BLEU and CIDEr scores remain very close to the baseline values, showing that improvements in token-level cross-entropy do not necessarily translate into stronger sequence-level metrics. In contrast, applying beam search to the optimised decoder yields consistent gains across all metrics: BLEU-4 increases from 0.1596 to 0.1852, METEOR rises from 0.1917 to 0.1941, and CIDEr improves from 0.4084 to 0.4635. These results highlight that decoding strategy has a substantial impact on caption quality—often greater than architectural or hyperparameter changes—aligning with prior observations in image captioning work [11].

5.3 Qualitative Examples and Attention Heatmaps

Qualitative examples help interpret these metrics. For many dog- and people-centric scenes, the baseline decoder tends to produce short, generic captions such as “a dog running on the beach”, whereas the beam-search model often adds modest but accurate detail, e.g. “a

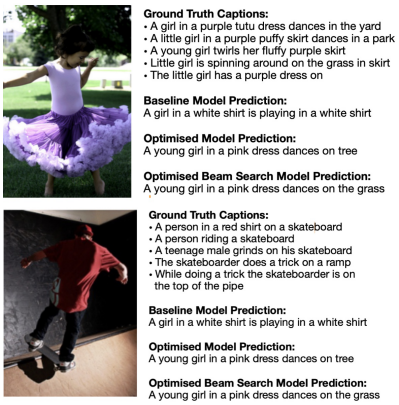


Figure 2: Qualitative examples comparing human references with captions from the baseline, optimised, and beam-search decoders.

black dog runs along the ocean surf”. On more complex scenes, both models still hallucinate attributes (for example mentioning a “man” when none is visible), reflecting dataset bias and the limited training size. Overall, Figure 2 shows several cases where the beam-search caption is closest to at least one human reference in both wording and semantics.

Figure 3: Attention heatmap for a dog-on-beach image. Left: original image. Right: averaged ViT attention overlay.

To improve interpretability, a separate ViT-B/16 model with `output_attentions=True` was used to compute attention heatmaps on selected test images. For each image, attention from the class token to patch tokens was averaged across layers and heads, reshaped to 14×14, and upsampled as a heatmap over the original image. Because the captioning encoder is frozen and uses the same pretrained checkpoint, these attention weights are effectively identical to those used in the main system.

In Figure 3, the human references describe a single black dog running along the surf, whereas the model predicts “two dogs are running in the surf”. The heatmap focuses on the dog and surf regions, showing that the encoder attends to the right areas while still making an object-counting error. This illustrates how the maps reveal what the visual backbone looks at, but not every token-level decision of the decoder.

5.4 Manual Error Taxonomy

Table 2: Manual error analysis of 30 failed captions.

Error type	Freq.	Example
Object hallucination	8	“A cat sits on the sofa” (no cat present).
Attribute error	9	“A boy in a red shirt” (shirt is clearly blue).
Relation error	5	“Man riding a horse” (man standing beside it).
Gender / bias	4	Long-haired boy described as “a girl”.
Repetition	2	“A dog runs runs in the grass.”

A small manual error analysis on 30 randomly selected test images with unsatisfactory predictions was conducted, and Table 2 summarises the main failure categories. Attribute errors and object hallucinations are the most frequent, followed by relation mistakes and occasional gender bias. These errors are consistent with dataset bias towards common objects and the limited supervision from training on a single caption per image.

6 Discussion and Critical Review

6.1 Coherent Summary

The project implemented an end-to-end ViT+Transformer image captioning pipeline on the Flickr8k dataset, covering EDA, preprocessing, baseline training, hyperparameter optimisation, and decoding analysis. A frozen pretrained ViT-B/16 encoder provided stable visual features, while a Transformer decoder was trained from scratch using cross-entropy loss. Despite training on only one caption per image, the models produced competitive BLEU, METEOR, and CIDEr scores for a dataset of this size. Optuna-based tuning alone resulted in limited improvements under greedy decoding, but combining the tuned decoder with beam search delivered clear and consistent gains: BLEU-4 increased from 0.1671 (baseline greedy) to 0.1852 (optimised beam), a 10.8% relative improvement; METEOR rose from 0.1821 to 0.1941 (a 6.6% gain); and CIDEr improved from 0.4165 to 0.4635 (an 11.3% gain). Compared with the same optimised decoder under greedy decoding, beam search also improved BLEU-4 from 0.1596 to 0.1852 (a 16.1% relative improvement) and METEOR from 0.1917 to 0.1941 (a 1.3% gain). Qualitative examples, the manual error taxonomy, and ViT-based attention heatmaps further support these trends by highlighting where the model captures key visual content and where it still reflects dataset biases.

6.2 Comparison with Previous Work and Justification

Compared with early CNN+LSTM captioners such as Show and Tell [15] and Show, Attend and Tell [16], our results remain competitive. Show, Attend and Tell reports a BLEU-4 of about 0.22 on Flickr8k, while our frozen-ViT system achieves 0.1852 using a much lighter decoder and without any encoder fine-tuning. This shows that a substantial portion of classic captioning performance can be achieved with a modern Transformer decoder and an effective decoding strategy.

Unlike Bottom-Up/Top-Down models that require object detectors [2], our approach stays detector-free and computationally lightweight, which aligns with coursework-level constraints. Although large vision–language models such as CLIP and BLIP [10, 8] achieve significantly higher scores on larger datasets, they rely on millions of image–text pairs and extensive pretraining. In contrast, this project emphasises transparency, reproducibility, and efficient use of limited data: a frozen ViT encoder appropriate for Flickr8k's scale, Optuna for sample-efficient tuning, and a beam width of $k = 3$, experimentally validated by consistent metric gains.

6.3 Critical Discussion

Several strengths and limitations are worth highlighting.

Strengths. *Data discipline.* Image-level splitting and the use of all five captions reduce leakage and follow good evaluation practice. BLEU, METEOR, and CIDEr are appropriate for this generative task.

Methodological clarity. The progression from a simple baseline to an Optuna-tuned model and then to beam search provides a clear improvement pathway, with decoding strategy having a larger impact on caption quality than architectural changes.

Stability. Freezing the encoder, using dropout, and applying early stopping prevent catastrophic overfitting, although a mild underfitting regime persists.

Interpretability. ViT attention heatmaps, qualitative examples, and the error taxonomy offer insight into model behaviour beyond scalar metrics.

Limitations. *Underfitting and frozen encoder.* Training and validation losses decrease together but remain relatively high, suggesting underfitting capped by the frozen encoder and limited data. Selectively unfreezing later ViT blocks with regularisation would be a natural extension.

Single-caption training. Using only one caption per image limits linguistic diversity; training on all five captions would better exploit the dataset.

Dataset bias. Flickr8k is biased towards people, dogs, and outdoor scenes, and the model inherits these biases, as reflected in the hallucination errors in Table 2.

Metric limitations. BLEU, METEOR, and CIDEr approximate human judgement but remain imperfect; the error taxonomy helps contextualise failure types.

Attention visualisation. The heatmaps come from a separate ViT instance with exposed attentions; they reliably show encoder focus but should be interpreted as indicative rather than fully causal explanations.

7 Conclusion and Future Work

This report showed that a frozen ViT encoder combined with a Transformer decoder can form an effective and interpretable image captioning system on the Flickr8k dataset. A simple baseline provided a reference point, Optuna explored decoder hyperparameters, and beam search decoding yielded consistent gains across BLEU, METEOR, and CIDEr (up to BLEU-4 = 0.1852, METEOR = 0.1941, and CIDEr = 0.4635) without changing the architecture. Qualitative examples, attention heatmaps, and a manual error taxonomy complemented these scores and highlighted where the model succeeds and fails.

Future work includes selectively unfreezing parts of the encoder with strong regularisation to reduce underfitting, training on all five captions per image to increase linguistic diversity, incorporating label smoothing or scheduled sampling to mitigate exposure bias, and evaluating the approach on larger datasets such as Flickr30k or MS-COCO. Exploring lightweight vision–language pretraining (e.g. CLIP-style encoders) within student compute budgets is another promising direction. Overall, the design choices—frozen encoder, modest decoder, Optuna tuning, and beam search—form a coherent and well-justified pipeline under the given constraints, supported by both quantitative metrics and qualitative analysis.

Acknowledgements

I would like to express my sincere gratitude to Dr. Mohammad Majid al-Rifaie and Dr. Hooman Oroojeni for their guidance and feedback throughout this project.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 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*, 2018.

[3] 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.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[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.

[6] Shadab Hussain. Flickr8k. `https://www.kaggle.com/datasets/shadabhussain/flickr8k`. Kaggle dataset, accessed 18 November 2025.

[7] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10):1–41, 2022.

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[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*, 2002.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[11] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. In *International Conference on Approximate Bayesian Inference*, 2022.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[14] 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*, 2015.

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015.