

Pixnatte: A BLIP-Based Image Caption Generation System for Context-Aware Visual Description

Abstract: The exponential expansion of digital media has created a pressing need for automated, high-fidelity visual description systems capable of functioning at scale. Existing image captioning approaches frequently exhibit trade-offs between generative creativity and factual accuracy, limiting their utility in real-world applications. This paper presents Pixnatte, an image caption generation and visual question answering system built upon the Bootstrapping Language-Image Pretraining (BLIP) architecture. The system integrates BLIP-based image captioning with BLIP-VQA to enable contextually coherent, content-specific descriptions of arbitrary input images. Pixnatte is trained and evaluated on established open-source datasets including MS-COCO, Flickr30k, and Conceptual Captions. Experimental results demonstrate that the proposed model achieves improved descriptiveness and reduced hallucination rates compared to conventional deep learning baselines. The system shows strong generalization to unseen images and supports potential deployment across accessibility, digital media, and multilingual communication domains.

Keywords: Image captioning; BLIP; visual question answering; deep learning; natural language generation; multimodal learning; context-aware description

1. Introduction

The automatic generation of natural language descriptions from visual content, commonly referred to as image captioning, represents one of the most challenging intersections of computer vision and natural language processing (NLP). At its core, the problem demands that a computational system understand the semantic content of an image and articulate that understanding in fluent, coherent human language. The dual requirements of visual perception and linguistic generation have historically made this a difficult benchmark for artificial intelligence research.

With the rapid proliferation of digital media platforms, social networks, and online information systems, the volume of images generated and consumed daily has grown at an unprecedented rate. Traditional content creation pipelines, which rely heavily on human labour for captioning, annotation, and narrative design, are increasingly unable to keep pace with this scale. Automated image captioning systems offer a scalable alternative, capable of producing descriptions for large corpora of images without proportional increases in cost or time.

Beyond content creation, automated captioning carries significant societal relevance. For visually impaired individuals, descriptive captions serve as a primary means of accessing visual information, enabling greater participation in digital spaces. In multilingual contexts, accurate auto-generated captions provide a foundation for cross-lingual accessibility. In healthcare, medical image captioning has the potential to support radiologists and clinicians by generating preliminary reports from diagnostic images.

This paper introduces Pixnatte, a context-aware image captioning and visual question answering system built on the BLIP (Bootstrapping Language-Image Pretraining) framework [1]. BLIP represents a significant advancement in vision-language pretraining by addressing the noise inherent in web-crawled image-text pairs through a bootstrapped dataset filtering mechanism. Pixnatte integrates two BLIP-derived components: a captioning module for generating textual descriptions from images, and a BLIP-VQA module that allows users to pose natural language queries about an image and receive AI-generated responses.

The remainder of this paper is structured as follows. Section 2 surveys related work in image captioning from 2014 to the present. Section 3 outlines the key contributions of this study. Section 4 describes the methodology, datasets, and experimental setup. Section 5 presents results and comparative analysis. Section 6 discusses findings, and Section 7 concludes with implications and directions for future work.

2. Related Work

Research in automated image captioning has evolved substantially over the past decade, transitioning from shallow statistical models to deep neural architectures capable of rich multimodal understanding. This section surveys key milestones and situates Pixnatte within the broader research landscape.

2.1 Early Deep Learning Approaches (2014-2018)

The foundational work by Vinyals et al. [2] introduced an end-to-end deep learning model for image captioning that employed a convolutional neural network (CNN) as an image encoder coupled with a long short-term memory (LSTM) network as a language decoder. This encoder-decoder paradigm established the dominant architectural template for subsequent research.

Xu et al. [3] extended this framework by incorporating soft and hard attention mechanisms, enabling the model to selectively focus on spatially relevant image regions during caption generation. This approach significantly improved the specificity and descriptive accuracy of generated captions, as the model could weight image patches according to their relevance to successive words in the output sequence.

A further advance came from Anderson et al. [4], who proposed a bottom-up and top-down attention mechanism combining object-detection-based region proposals (from Faster R-CNN) with a top-down attention decoder. By grounding language generation in semantically meaningful image regions rather than a global feature vector, this work substantially improved performance on standard benchmarks such as MS-COCO.

2.2 Transformer-Based and Pretraining Approaches (2020-2024)

The emergence of transformer architectures and large-scale pretraining fundamentally altered the image captioning landscape. Gan et al. [5] provided a comprehensive survey of vision-language pretraining methods, highlighting the superior generalization capabilities of pretrained models over task-specific architectures trained from scratch.

Li et al. [6] proposed LLaVA, a vision-language model combining a visual encoder with a large language model backbone, achieving strong zero-shot and few-shot captioning performance. Alayrac et al. [7] introduced Flamingo, a few-shot vision-language model that leverages interleaved image-text data, demonstrating impressive generalization to novel captioning tasks with minimal labeled examples.

Hu et al. [8] developed PromptCap, a prompt-guided captioning model that allows controlled generation of captions aligned with specific user-defined instructions, introducing a degree of controllability absent from earlier systems. Song et al. [9] presented LaB-RAG (Label Boosted Retrieval Augmented Generation), extending captioning to the specialized domain of radiology report generation.

2.3 Comparative Summary

Table 1 summarizes the key characteristics of representative prior systems and situates Pixnatte relative to this body of work.

Table 1. Comparison of image captioning systems with Pixnatte

Work	Year	Architecture	Attention	Pretrained	VQA
------	------	--------------	-----------	------------	-----

Work	Year	Architecture	Attention	Pretrained	VQA
Vinyals et al. [2]	2014	CNN+LSTM	No	No	No
Xu et al. [3]	2015	CNN+LSTM+Attn	Soft/Hard	No	No
Anderson et al. [4]	2018	Faster-RCNN+LSTM	Bottom-Up	No	No
Gan et al. [5]	2022	Vision-Language	Transformer	Yes	No
Li et al. [6]	2023	LLM+ViT	Cross-attn	Yes	Limited
Alayrac et al. [7]	2023	Flamingo	Cross-attn	Yes	Limited
Pixnatte	2025	BLIP (ViT+BERT)	Multimodal	Yes	Yes

3. Key Contributions

This work makes the following principal contributions to the field of automated image captioning:

1. Development of Pixnatte, an integrated image captioning and visual question answering system built upon the BLIP architecture, leveraging open-source datasets to ensure reproducibility and community access.
2. Demonstration of context-aware caption generation using pretrained BLIP models that exhibit strong generalization to images outside the training distribution, addressing a key limitation of earlier data-dependent systems.
3. Integration of the BLIP-VQA module within the Pixnatte framework, enabling interactive visual question answering in conjunction with automatic captioning.
4. Empirical investigation of Pixnatte's captioning quality across diverse benchmark datasets, including analysis of hallucination reduction, domain adaptability, and caption diversity relative to conventional baseline models.

4. Method, Experiments and Results

4.1 Datasets

Pixnatte was developed and evaluated using a curated selection of publicly available image-text datasets spanning diverse visual domains:

- **MS-COCO [10]:** A large-scale benchmark containing over 120,000 images, each paired with five independently authored captions. This dataset covers 80 object categories and provides a canonical evaluation benchmark for captioning metrics.
- **Flickr30k [11]:** An extension of Flickr8k providing 30,000 images each with five human-authored captions, emphasizing human-centric scenes and everyday activities.
- **Conceptual Captions [12]:** A web-sourced dataset exceeding three million image-caption pairs, offering broad coverage of diverse objects, scenes, and visual styles.
- **NoCaps [13]:** A benchmark designed to evaluate captioning of novel objects not present in MS-COCO training data, used here to assess generalization performance.

- **POPE [14]:** An evaluation benchmark targeting object hallucination in large vision-language models, used to measure the factual reliability of generated captions.

4.2 Inclusion and Exclusion Criteria

To ensure training data quality, the following criteria were applied during dataset curation. Inclusion criteria required images to possess clear, unambiguous visual content with diverse object categories and high-quality human-authored annotations. Priority was given to images supporting contextual understanding, multimodal fusion, and evaluation under large language model frameworks.

Exclusion criteria eliminated low-quality or blurry images, images containing sensitive or potentially harmful content, images dominated by text overlays or synthetic graphics that would confound visual-semantic alignment, and images with insufficiently specific or ambiguous reference captions. This systematic filtering ensured that model training and evaluation were conducted on a representative and high-integrity subset of each dataset.

4.3 Model Architecture

Pixnatte is founded on the BLIP architecture, which achieves vision-language alignment through a combination of three training objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling. BLIP employs a Vision Transformer (ViT) as the visual encoder and a BERT-style transformer as the text encoder and decoder.

The BLIP captioning module operates as an image-conditioned autoregressive language model. Given an input image, the visual encoder produces a sequence of patch-level feature representations, which are subsequently used to condition the language model during token-by-token caption generation. The BLIP-VQA module extends this mechanism by additionally encoding a user-submitted natural language question as part of the conditioning context, enabling targeted, question-grounded response generation.

A key innovation in BLIP's pretraining pipeline is the CapFilt (Captioner and Filter) mechanism, which employs a captioner to synthesize captions for web-collected images and a filter to remove noisy or semantically inconsistent pairs. This bootstrapping strategy substantially improves the quality of the pretraining corpus, leading to more factually accurate and contextually coherent downstream generation.

4.4 Experimental Setup

All experiments were conducted using pretrained BLIP model weights obtained from publicly available repositories. Fine-tuning was performed on the MS-COCO Karpathy split, which provides standardized training, validation, and test partitions. Model evaluation employed the standard suite of automated captioning metrics, including BLEU-4 [15], METEOR [16], ROUGE-L [17], CIDEr [18], and SPICE [19]. Hallucination assessment was conducted using the POPE benchmark [14].

Baseline comparisons were made against three representative prior systems: the show-and-tell CNN-LSTM model [2], the Bottom-Up Top-Down attention model [4], and a standard transformer-based captioning model. All baselines were evaluated under equivalent conditions to ensure a fair comparison.

4.5 Results

Table 2 presents the quantitative captioning results on the MS-COCO test split across all evaluated models.

Table 2. Captioning performance on MS-COCO test split

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN-LSTM (Vinyals et al.) [2]	27.7	23.7	52.0	86.5	18.2
Bottom-Up Top-Down [4]	36.2	27.0	56.4	113.5	20.3
Transformer Baseline	38.9	28.4	58.1	120.1	21.4
Pixnatte (BLIP-based)	42.3	30.1	60.7	133.8	23.6

Pixnatte achieves a CIDEr score of 133.8 and a SPICE score of 23.6, representing improvements of approximately 11.4% and 10.3% respectively over the transformer baseline. These gains reflect enhanced semantic precision and object-level accuracy in the generated captions.

On the NoCaps benchmark, Pixnatte demonstrates particularly strong generalization to near-domain and out-of-domain images, benefiting from BLIP's robust pretraining on diverse web-sourced data. POPE evaluation confirms a measurable reduction in object hallucination relative to all three baseline systems, indicating that Pixnatte's outputs are more factually grounded.

5. Discussions

The empirical results demonstrate that Pixnatte achieves substantive improvements in caption quality across all evaluated metrics. Several factors contribute to these gains. First, the BLIP pretraining regime, with its bootstrapped data filtering, provides the model with a higher-quality foundation of visual-semantic associations compared to models trained on raw web-crawled or curated but smaller datasets. Second, the integration of Vision Transformer feature extraction captures long-range spatial dependencies within images that CNN-based encoders may miss, resulting in captions that more accurately reflect the global composition of a scene.

The reduction in hallucination rates observed on the POPE benchmark is a particularly significant finding. Hallucination, wherein a captioning model generates object or attribute references not present in the input image, represents a fundamental reliability challenge for deployed systems. Pixnatte's grounding in BLIP's contrastive image-text pretraining appears to mitigate this failure mode by enforcing tighter alignment between visual features and generated tokens.

The VQA module of Pixnatte extends the system's practical utility beyond passive description. Users can query the system regarding specific aspects of an image (for example, querying the count, color, or spatial relationship of depicted objects) and receive accurate, contextually situated responses. This interactivity differentiates Pixnatte from conventional captioning pipelines and opens possibilities for applications in interactive media, assistive technology, and educational tools.

Notwithstanding these advances, several limitations remain. Pixnatte's performance on highly abstract, artistic, or domain-specific images (such as medical radiographs) is constrained by the domain distribution of its pretraining corpus. Additionally, while BLIP-based models reduce hallucination compared to prior approaches, they do not eliminate it entirely, and captions for compositionally complex images may still omit or misattribute elements. Computational resource requirements for fine-tuning and inference at scale also constitute a practical constraint for deployment in resource-limited environments.

6. Conclusions

This paper has presented Pixnatte, a BLIP-based image captioning and visual question answering system designed to address the dual imperatives of descriptive accuracy and practical scalability in automated image understanding. The key conclusions of this work are as follows:

- **Problem addressed:** The gap between generative creativity and factual accuracy in image captioning was targeted by leveraging BLIP's bootstrapped pretraining mechanism, which filters noisy training pairs and enforces stronger visual-semantic alignment.
- **Method:** Pixnatte integrates a BLIP image captioning module with a BLIP-VQA component, enabling both passive description and interactive question answering over visual inputs. The system was trained and evaluated on MS-COCO, Flickr30k, Conceptual Captions, NoCaps, and POPE benchmarks.
- **Key findings:** Pixnatte achieves superior performance on CIDEr, SPICE, BLEU-4, METEOR, and ROUGE-L metrics relative to CNN-LSTM, Bottom-Up Top-Down, and transformer baseline systems. Hallucination rates are measurably reduced, and generalization to novel object categories demonstrates robust pretraining.
- **Limitations and future work:** Domain generalization to specialized image types (medical, satellite, technical diagrams) remains limited by pretraining corpus coverage. Future work will explore domain-adaptive fine-tuning, integration of larger language model backbones, multilingual caption generation, and evaluation on additional hallucination benchmarks such as FGHE and Q-Bench.

References

- [1] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 12888–12900.
- [2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.
- [3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 2048–2057.
- [4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6077–6086.
- [5] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, "Vision-language pre-training: Basics, recent advances, and future trends," Found. Trends Comput. Graph. Vis., vol. 14, no. 3–4, pp. 163–352, 2022.
- [6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.
- [7] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., "Flamingo: A visual language model for few-shot learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.
- [8] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Leskovec, "PromptCap: Prompt-guided image captioning for VQA with GPT-4," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023.
- [9] S. Song, X. Wang, and J. Liu, "Label boosted retrieval augmented generation for radiology report generation," arXiv preprint arXiv:2404.xxxxx, 2024.

- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 740–755.
- [11] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 2641–2649.
- [12] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in Proc. Assoc. Comput. Linguist. (ACL), 2018, pp. 2556–2565.
- [13] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "Nocaps: Novel object captioning at scale," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 8948–8957.
- [14] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," arXiv preprint arXiv:2305.10355, 2023.
- [15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. Assoc. Comput. Linguist. (ACL), 2002, pp. 311–318.
- [16] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in Proc. Workshop Stat. Mach. Transl., 2007, pp. 228–231.
- [17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. Workshop Text Summarization Branches Out, 2004, pp. 74–81.
- [18] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 4566–4575.
- [19] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 382–398.