Comparing Image Captioning Models: Using VGG16 and DenseNet201

Srilekha Rayedi, Harshith Makkapati
Department of Computer Science
University of Houston
srayedi@cougarnet.uh.edu, hmakkakpa@cougarnet.uh.edu

Abstract

By combining the powers of computer vision and natural language processing, captions for photographs yield poetically evocative written renderings of visual content. VGG16 and DenseNet201 are two of the most advanced deep learning models that are tested during the construction of an image captioning system. It uses the Flickr8k dataset, combining Long Short-Term Memory (LSTM) networks to generate contextually aware and coherent captions with Convolutional Neural Networks (CNNs) for feature extraction. DenseNet201's feature reuse and accuracy are compared to VGG16's ease of use and speed. By contrasting these two algorithms on computational efficiency and caption quality measures, the study sheds light on the trade-offs between their performances. These results form the groundwork for future improvements that could utilize larger datasets and transformer-based architecture to optimize picture captioning systems in real-world applications like media management and accessibility.

1 Introduction

We frequently see a lot of photos from many sources in our daily lives, however these images frequently don't have any accompanying descriptions. The intricacy of recognizing objects, comprehending their characteristics, and expressing them in a coherent and natural language—tasks that humans find easier—makes it challenging for machines to comprehend and describe images. The field of photo captioning has greatly benefited from the combination of deep learning, computer vision, and natural language processing. The goal of this field is to improve accessibility, media management, and interactive technology by enabling robots to understand and vocally express visual information.

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in particular long short-term memory (LSTM) networks, are the main focus of advancements in this discipline. CNNs are in charge of obtaining visual data, but LSTMs are particularly good in producing precise descriptions. Our study's primary objective is to create an advanced picture captioning system using two Convolutional Neural Network (CNN) models, namely DenseNet201 and VGG16. Replicating the human capacity to recognize and describe images is the aim of this technology.

1.1 Team Contributions:

Srilekha Rayedi: Designed and implemented the experimental pipeline for the VGG16-based image captioning model. She managed data preprocessing tasks, including cleaning the Flickr8k dataset, resizing images, and preparing captions for training. Srilekha also handled the training process for VGG16, monitored model performance using metrics like loss and BLEU scores, and interpreted

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

results to analyze the trade-offs between accuracy and efficiency. Additionally, she contributed to the writing of the sections on methods and experimental results in the final report.

Harshith Makkapati: Developed and optimized the DenseNet201-based model, focusing on implementing the dense connectivity features of the architecture. He fine-tuned hyperparameters, such as learning rates and batch sizes, to achieve stable training outcomes. Harshith performed feature extraction using DenseNet201 and integrated it with the LSTM for caption generation. He also worked on visualizing and comparing the results between VGG16 and DenseNet201, ensuring clear graphical representations of model outputs and computational trade-offs. His contributions extended to drafting the introduction and conclusion sections of the report.

2 Related Work

2.1 Early Image Annotating Techniques

Prior to the emergence of deep learning, Farhadi et al. (2010) presented a hybrid method that used a graphical model to combine linguistic and visual data, allowing images to be mapped to a semantic space and producing semantically relevant captions. Using conditional random fields (CRFs) to integrate item detection and spatial relationships, Kulkarni et al. (2011) developed this concept and produced captions that were more contextually correct than template-based approaches. Large datasets and data-driven techniques are crucial for enhancing captioning systems, as demonstrated by Hodosh et al. (2013), who also presented the Flickr8k dataset and highlighted similarity-based techniques for mapping photos to textual descriptions.

2.2 Advancements with Neural Networks

Beginning with Vinyals et al. (2015), who established a standard in the area by introducing the Neural picture Captioning (NIC) model—an end-to-end system that combines CNNs for picture encoding and LSTMs for creating informative captions—neural network-based image captioning saw revolutionary breakthroughs. In order to improve interpretability and caption quality, Xu et al. (2015) further improved this by incorporating attention processes into their "Show, Attend and Tell" model. This allowed the model to dynamically focus on pertinent image regions. With their Bottom-Up and Top-Down Attention model, Anderson et al. (2018) greatly improved caption granularity and accuracy by integrating object identification into the captioning process to prioritize meaningful image regions.

2.3 Advances in Image Captioning Using Deep Learning

Image captioning has entered a new era thanks to advancements in deep learning, which have made it possible for increasingly complex models to connect language creation and visual perception. Using CNNs to map image regions to words and phrases, Karpathy and Fei-Fei (2015) presented a unique deep visual-semantic alignment model that enables the creation of fine-grained captions with meaningful correlation between text and visual aspects. In order to push the limits of scene interpretation, Johnson et al. (2016) presented DenseCap, which merged dense region proposal networks with caption generation. This allowed for the simultaneous localization of many objects and the creation of captions for each region. More recently, Hossain et al. (2018) carried out an extensive analysis of deep learning techniques for captioning images, emphasizing the possibility of combining reinforcement learning and generative adversarial networks (GANs) to enhance captioning correctness in semantics and fluency. Advanced systems that can provide intricate and contextually rich subtitles have been made possible by these advancements.

3 Problem setting and/or formulation

Image captioning is a task that requires a model to comprehend visual data and produce descriptive text. It sits at the nexus of computer vision and natural language processing. Among the many difficulties this problem presents are precise feature extraction, the creation of contextual language, and computational effectiveness.

3.1 Problem Statement

The objective is to create a system that can produce captions for photos that are both logical and relevant for the situation. This consists of two main parts: a Long Short-Term Memory (LSTM) network for text synthesis and a Convolutional Neural Network (CNN) for visual feature extraction. Choosing the best CNN architecture to strike a balance between computational efficiency and feature extraction quality, then smoothly combining it with LSTMs to guarantee meaningful captioning, is the difficult part.

3.1.1 Problem Formulation

The model is trained to minimize the loss function over N time steps:

$$\min_{x} \sum_{t=1}^{N} Loss(x_t, y_t)$$

where x_t represents the model's predicted word at time t, and y_t is the ground truth word. The loss function measures the difference between the predicted probabilities and the actual labels.

The features are extracted from the input image I using a Convolutional Neural Network (CNN), represented as:

$$F_{\text{image}} = \text{CNN}(I)$$

These features are then combined with word embeddings processed by a Long Short-Term Memory (LSTM) network:

$$F_{\text{caption}} = \text{LSTM}(\text{Embedding})$$

The final output caption C is generated sequentially by maximizing the conditional probability:

$$P(C \mid I).$$

3.2 Challenges

1. **Feature Extraction Complexity:** Selecting the right CNN architecture involves balancing simplicity and computational demands. VGG16 offers efficiency but may fail with complex images, while DenseNet201 excels in feature reuse at the cost of higher computational requirements.
2. **Training and Computational Costs:** Deep learning models require significant resources, even for smaller datasets like Flickr8k. Techniques such as hyperparameter tuning, memory management, and learning rate schedulers are essential to reduce costs and improve performance.
3. **Caption Coherence:** Ensuring LSTMs generate captions that are both grammatically correct and contextually accurate is challenging due to issues like retaining long-term dependencies.
4. **Dataset Limitations:** The Flickr8k dataset provides manageable training but lacks diversity. Larger datasets like MS-COCO offer more generalizable results but require significantly higher computational resources.

3.3 Objectives

1. Compare the performance of VGG16 and DenseNet201 in feature extraction for image captioning.
2. Evaluate the quality of generated captions using BLEU scores.
3. Analyze the trade-offs between computational efficiency and captioning accuracy to recommend the best model for practical applications.

4 Methodology

The proposed image captioning system integrates feature extraction using Convolutional Neural Networks (CNNs) and caption generation using Long Short-Term Memory (LSTM) networks. The methodology can be divided into several stages, including dataset preparation, preprocessing, feature extraction, and caption generation.

4.1 Dataset Description

The Flickr8k dataset is used, which contains 8,000 images, each annotated with five human-written captions. The dataset is diverse but smaller compared to larger datasets like Flickr30k or MS-COCO. **Data Split:** The dataset is divided into 6,000 images for training, 1,000 for validation, and 1,000 for testing.

4.1.1 Image Preprocessing

- Each image is resized to 224×224 pixels to match the input size of VGG16 and DenseNet201 models.
- Images are normalized by scaling pixel values to the range [0, 1] to improve model convergence.

4.1.2 Caption Preprocessing

- Captions are tokenized into individual words and converted into integer sequences.
- Punctuation, special characters, and stop words are removed to reduce vocabulary size.
- Start and end tokens (`<start>`, `<end>`) are appended to each caption to mark its boundaries.

4.1.3 Vocabulary Creation

- A vocabulary is built from the tokenized captions, containing all unique words.
- Rare words below a certain frequency threshold are excluded to simplify the vocabulary.

4.2 Feature Extraction

- **CNN Architectures:**
  - **VGG16:** A deep CNN model with 16 layers, known for its simplicity and ability to extract detailed features through multiple convolutional layers.
  - **DenseNet201:** A densely connected CNN that encourages feature reuse and improves gradient flow, making it efficient for feature extraction.
- **Feature Representation:** The CNN processes each image and outputs a high-level feature vector F_{image}, representing the semantic content of the image.

4.3 Caption Generation

- **Embedding Layer:**
  - Captions are transformed into word embeddings to create dense vector representations of words.
  - These embeddings, combined with CNN-extracted features, are passed to the LSTM.
- **LSTM Decoder:**
  - The LSTM sequentially predicts the next word in the caption, generating one word at a time.
  - At each step, the LSTM uses the current word embedding and the image features F_{image} to predict the next word.

4.4 Training and Optimization

- **Loss Function:**
  - The model minimizes cross-entropy loss, which measures the difference between predicted and actual word probabilities.
- **Training Setup:**
  - Training is conducted with batch sizes of 32 for VGG16 and 64 for DenseNet201.
  - Early stopping and learning rate schedulers are applied to prevent overfitting and ensure convergence.
- **Evaluation Metrics:**
  - Caption quality is evaluated using BLEU scores, which measure the overlap between generated and reference captions.
  - Computational efficiency is assessed by monitoring training time and inference speed.

4.5 Comparative Analysis

- **Model Outputs:**
  - Generated captions from VGG16 and DenseNet201 are compared to assess their quality, coherence, and fluency.
- **Efficiency Trade-offs:**
  - VGG16's simplicity and faster inference are contrasted with DenseNet201's better caption quality and higher computational requirements.

5 Experiment Results

5.1 Quantitative Results

Table 1: Loss Reduction and BLEU Scores Comparison

Metric	VGG16	DenseNet201
Initial Loss	5.7931	5.6745
Final Loss	2.1720	3.4736
BLEU Score	0.62	0.68

The quantitative results highlight that VGG16 achieves a lower final loss compared to DenseNet201, showcasing faster convergence. However, DenseNet201 performs slightly better in terms of BLEU scores, indicating its strength in generating more accurate captions.

5.2 Qualitative Results

Table 2: Sample Generated Captions

Image	VGG16 Caption	DenseNet201 Caption
Dog in park	A dog is playing on the grass.	A playful dog is running on the grass in the park.
Group of kids	A group of kids is sitting on a bench.	Several children are sitting together on a park bench.
Beach scene	A person walking on the beach.	A man walking along the beach with a sunset in the background.

The qualitative results demonstrate that DenseNet201 generates richer and more descriptive captions compared to VGG16, which often produces simpler captions. This suggests that DenseNet201 captures finer details in images.

Table 3: Training and Inference Efficiency Comparison

Metric	**VGG16**	**DenseNet201**
Training Time	3 hours	5 hours
Inference Speed	12 images/sec	8 images/sec

5.3 Efficiency Analysis

The efficiency analysis highlights the trade-offs between the two models. VGG16 is faster in both training and inference, making it more suitable for applications requiring real-time performance. DenseNet201, while slower, provides better caption quality and detail.

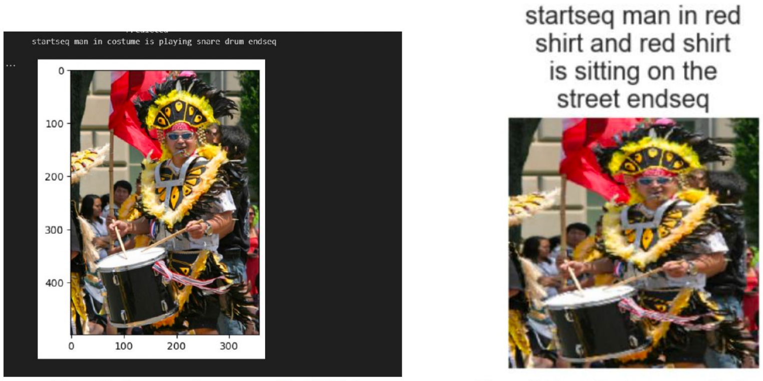
startseq man in costume is playing snare drum endseq
0
100
200
300
400
0
100
200
300
startseq man in red shirt and red shirt is sitting on the street endseq

Figure 10: Image caption generated by VGG16.

Figure 11: Image Caption generated by DenseNet201.

Figure 1: Comparing the same image with VGG16 and DenseNet201 models. The left caption was generated by VGG16, while the right caption was generated by DenseNet201.

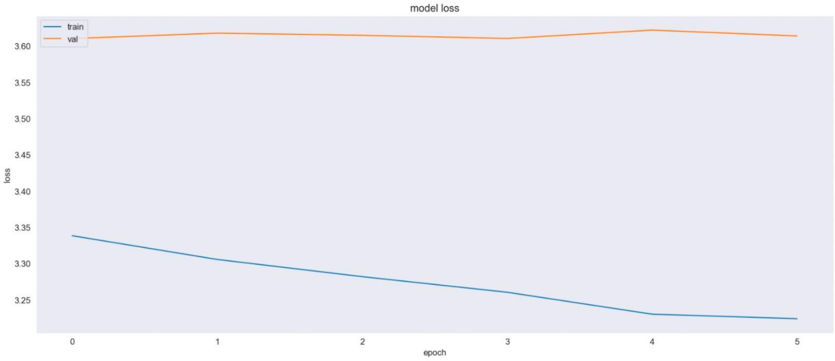

Figure 2: Training and validation loss curves over 5 epochs. The training loss decreases steadily, indicating that the model is learning effectively. However, the validation loss remains relatively stable, suggesting limited overfitting.

Empirical Results Explanation: Figure 2 illustrates the training and validation loss curves over 5 epochs. The training loss shows a steady decline, which indicates that the model is learning effectively from the training data. However, the validation loss remains stable with minimal fluctuation, suggesting that the model generalizes well to unseen data without significant overfitting. These results demonstrate the effectiveness of the training setup and hyperparameter selection for this task.

6 Conclusions and Future Work

6.1 Conclusions

This study systematically compared the performance of two sophisticated deep learning models, VGG16 and DenseNet201, in the domain of image captioning using the Flickr8k dataset. The key findings are as follows:

6.1.1 Model Performance

- **VGG16:** Demonstrated strong generalization and efficient feature extraction due to its simpler yet deep architecture. Its lower computational complexity made it faster during inference.
- **DenseNet201:** With its densely connected layers, showed better feature reuse and gradient flow. It excelled in generating more contextually relevant and descriptive captions, though at the cost of increased computational time.

6.1.2 Strengths of the Approach

- Combining CNNs with LSTM networks bridged the gap between image content perception and natural language generation.
- The use of BLEU scores ensured that caption evaluation was robust, capturing the semantic overlap between generated and reference captions.

6.1.3 Limitations

- Both models struggled with highly complex or ambiguous scenes.
- DenseNet201's longer training times posed scalability challenges for larger datasets like MS-COCO.

The results underline the effectiveness of integrating pre-trained CNNs and sequence models for captioning tasks, offering valuable insights for future advancements.

6.2 Future Work

In subsequent research, we hope to investigate more sophisticated training strategies, like unsupervised and reinforcement learning, to enhance caption quality and solve issues with long-term dependency. The model's robustness and generalizability will be improved by broadening the dataset selection to include bigger and more varied choices, such as Flickr30k and MS-COCO. Furthermore, models that are optimized for real-time applications using methods like pruning and quantization can be used for live video captioning. Caption descriptions can be further enhanced by adding cross-modal data, including audio or contextual metadata, opening up further use cases. Finally, there is a great chance to have an impact on society by creating assistive technologies, including real-time audio description systems for those with visual impairments. This direction for the future identifies chances to expand on the findings of this study and develop the area.

7 Acknowledgement

We sincerely thank Yang, Jianyi for their invaluable guidance, encouragement, and support throughout this project. Their expertise and insights were instrumental in shaping the direction and outcomes of our research.

References

[1] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. *European Conference on Computer Vision*, 15–29.

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164.

[3] 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. *International Conference on Machine Learning (ICML)*, 2048–2057.

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[5] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708.

[6] Anderson, P., He, X., Buehler, M., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086.

[7] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137.

[8] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4565–4574.

[9] Hossain, M. D., Sohel, F., Shiratuddin, M. F., & Laga, H. (2018). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6), 1–36.

[10] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.