iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77386

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

Comparative Analysis of Pretrained Models in Prefix-Based Image Captioning Tasks under Limited Training Conditions

Sumedha Arya

Abstract: ***Image captioning task requires effective combination of visual feature extraction and natural language generation. This study compares four pre-trained models such as Vision Transformer (ViT-B/16), ResNet-18, VGG-16, and DenseNet-121 when applied as frozen feature extractors in a prefix-based captioning framework using a partially trainable BERT-base-uncased text encoder. Experiments were conducted on a 32,000-image subset of the MS COCO 2017 captions dataset (28,000 training, 4,000 validation) under a limited training environment. Performance was evaluated using training cross-entropy loss. Results show DenseNet-121 achieved the lowest final loss (0.2894), followed by VGG-16 (0.3198), ResNet-18 (0.3935), and ViT-B/16 (0.7002). DenseNet-121 demonstrated superior feature richness and fastest generalization, while ViT exhibited slowest convergence. These findings suggest that, under resource-constrained scenarios with frozen backbones, DenseNet-121 is the most effective choice among the evaluated architectures.***

Keywords: Image Captioning, Prefix-Based Captioning, Frozen Vision Backbone, DenseNet 121, Vision Transformer.

I. INTRODUCTION

A single image may also contain information that can be in a large amount. Every day, a huge number of images are generated on social media platforms and various other sources. Deep learning makes it possible to automatically label images, which can reduce or even eliminate the need for human involvement in this task [1]. This helps save time and effort.

One of the main challenges today is the large volume of images and related text available online. Another challenge is that image data changes frequently and often contain noise. Because of this, most raw data cannot be directly used for image captioning models and must be cleaned before use [2]. In general, image captioning is defined as an automatic generation of a meaningful sentence to describe what is happening in an image. To train an image captioning model, a large dataset with accurately labeled images is required. Image captioning is a difficult task because the model must first understand the image and then generate a correct sentence. It needs to extract useful features from the image and convert those features into natural language.

Although image captioning has improved a lot, still large computational resources are required, along with large datasets, and long training times. Also, there is still limited understanding of how well different pre-trained image models perform when they are kept completely frozen and only small additional layers are trained under very limited training conditions.

To address this, our work focusses on four popular pre-trained vision models—ViT-B/16, ResNet-18, VGG-16, and DenseNet-121—by using them as frozen feature extractors in a simple prefix-based image captioning model. The image features are converted into a single prefix token and passed to a partially trainable BERT text encoder. The main goal of this study is to find out which frozen vision model learns fastest and achieves the lowest training loss under strict limitations. The detailed results and their implications are discussed in the upcoming sections.

II. REVIEW OF THE LITERATURE

Most existing image captioning techniques are dependent on deep learning models such as Recurrent Neural Networks (RNNs) trained using Maximum Likelihood Estimation (MLE). They are used to generate descriptive captions for images. Although MLE improves training stability and accuracy, it suffers from a critical limitation known as exposure bias. In this case, the model trained on ground-truth sequences rely on its own predictions during inference. This mismatch often leads to degraded performance and captions that do not align well with human quality judgment [6].

To address these limitations, researchers have proposed Generative Adversarial Networks (GANs), as an alternative to MLE-based models. GANs produce visually realistic synthetic images that are difficult for humans to distinguish from real images [3]. In this framework, two deep learning based neural networks are used.

One is called a generator, that produces samples, while another is called a discriminator that evaluates their authenticity. Through adversarial training, both networks improve simultaneously, leading to more accurate and natural outputs [4]. This approach has been extended to image captioning, where the generator creates captions and the discriminator evaluates both the linguistic quality and consistency with the features of the image [7].

However, applying GANs to language generation still have certain issues. Unlike images, text consists of discrete tokens, which prevents direct gradient backpropagation from the discriminator to the generator. This makes conventional GAN training ineffective for natural language tasks [5]. To overcome this issue, many approaches adopt a Reinforcement Learning (RL) framework, where the caption generator is treated as an agent, and each generated word represents an action. The reward signal is provided by the discriminator, which allows for the estimation of the gradient using policy optimization techniques [6].

Despite its advantages, RL-based captioning also introduces a difficulty. There is a lack of intermediate rewards during sequence generation. In this process, the model receives feedback after the full caption is completed, making it difficult to evaluate the contribution of individual words [7]. To solve this problem, advanced methods employ policy gradient algorithms combined with Monte Carlo rollout strategies to estimate intermediate rewards. These techniques allow the model to consider long-term dependencies and improve the overall coherence and quality of the generated captions.

III.RESEARCH METHODOLOGY

The objective of this research is to utilize and compare four pre-trained vision models such as ViT-B/16, ResNet-18, VGG-16, and DenseNet-121—for an image captioning task. These models are used as frozen visual feature extractors within a prefix-based captioning framework. The comparison is performed based on limited training epochs to analyze which model adapts faster and performs better.

A. Research Design

This work follows an experimental and comparative research design, where only the model is varied while all other components remain constant to ensure a fair comparison. Key elements used in this process are:

1) Independent variable: Type of model
2) Dependent variable: Training loss (cross-entropy)
3) Controlled factors:

- Same dataset (COCO 2017)
- Same model architecture
- Same BERT encoder
- Same hyperparameters
- Same number of epochs (3)
- Same random seed (42)

B. Dataset and Preprocessing

The experiments use the MS COCO 2017 caption dataset. A total of 32,000 images were randomly selected, with 28,000 used for training and 4,000 used for validation. Caption preprocessing includes punctuation removal, conversion to lowercase, and normalization of extra spaces. During training, one random caption is selected per image.

C. Image Processing

All images are processed uniformly before being given to the model for training. The following steps were applied in this process:

1) Resize to 224 × 224
2) Conversion to tensor
3) Normalization using ImageNet mean and standard deviation
4) No data augmentation applied

D. Model Architecture

A prefix-based image captioning framework is applied. The image is first encoded using a frozen visual model. The extracted features are projected into the text embedding space and added as a prefix token to the input sequence of a BERT-based text encoder. The model then predicts the next token using a linear classification head.

E. Main Components

1) Frozen image encoder (ViT / ResNet / VGG / DenseNet)
2) BERT-base-uncased text encoder (partially trainable)
3) Image projection layer
4) Prefix fusion with text tokens
5) Linear prediction head for caption generation

F. Vision Backbones

Compared Four architectures were evaluated under identical experimental settings:

1) ViT-B/16 (Transformer-based 768-dimensional features)
2) ResNet-18 (512-dimensional features based on CNN)
3) VGG-16 (512-dimensional features based on CNN)
4) DenseNet-121 (1024-dimensional features based on CNN)

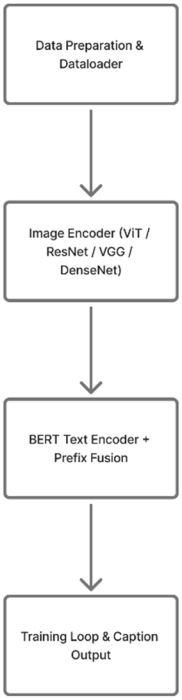


Fig 1. Model Architecture

All models were pre-trained on ImageNet and kept frozen during training.

G. Training Setup

All models were trained using identical hyperparameters to ensure correct comparison.

H. Training Configuration

1) Batch size: 24
2) Epochs: 3
3) Optimizer: AdamW
4) Learning rate: 4 × 10^{-5}
5) Loss function: Cross-entropy loss
6) Gradient clipping applied
7) Random seed fixed to 42 for reproducibility

I. Evaluation Strategy

Model performance is evaluated using training loss per epoch. Since the goal is comparative analysis under limited training conditions, the final loss after 3 epochs is used as the primary comparison metric. Lower loss indicates better learning and faster adaptation to the captioning task.

IV. RESULTS ANALYSIS

According to the research methodology, each model was trained for three epochs on the same dataset. The following table summarizes the training cross-entropy loss obtained across epochs, along with the relative performance ranking.

TABLE I
TRAINING LOSS COMPARISON AFTER 3 EPOCHS

Backbone	Epoch 1	Epoch 2	Epoch 3	Δ Loss (Ep1 → Ep3)	Iter/s	Rank
DenseNet-121	3.4126	0.5431	0.2894	-3.1232	7.3–7.4	1st
VGG-16	3.6245	0.6036	0.3198	-3.3047	6.4–6.9	2nd
ResNet-18	3.8717	0.7466	0.3935	-3.4782	7.6	3rd
ViT-B/16	5.0947	1.4750	0.7002	-4.3945	5.0–5.1	4th

DenseNet-121 achieved the lowest final training loss (0.2894), clearly outperforming all other models. This indicates that the dense connectivity pattern of DenseNet provides highly informative visual features. These features can be effectively utilized even when the model remains completely frozen and only the projection layers and text encoder components are trained.

Strong Performance of Classical CNN Architectures VGG-16, despite being an older architecture, achieved the second-best performance, outperforming both ResNet-18 and ViT-B/16. This suggests that in highly constrained training scenarios, strong and diverse feature representations may be more beneficial, rather than complex architectural advancements.

ViT-B/16 showed the weakest performance, with a relatively high final loss (0.7002) even after three epochs. The slower convergence may be due to further requirement of more fine-tuning and larger datasets to fully exploit their representational capacity. Additionally, the use of the [CLS] token as a prefix representation may be less suitable for direct alignment with a frozen text encoder.

ResNet-18 achieved the fastest training speed, approximately 7.6 iterations per second, while ViT-B/16 was the slowest, approximately 5 iterations per second. This represents an important practical consideration when computational resources or time are limited.

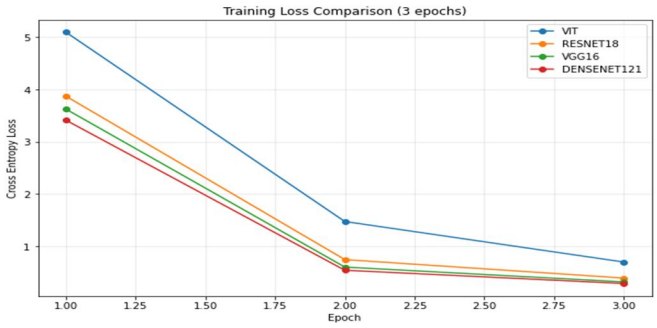


Fig 2. Training Loss Comparision

V. CONCLUSION

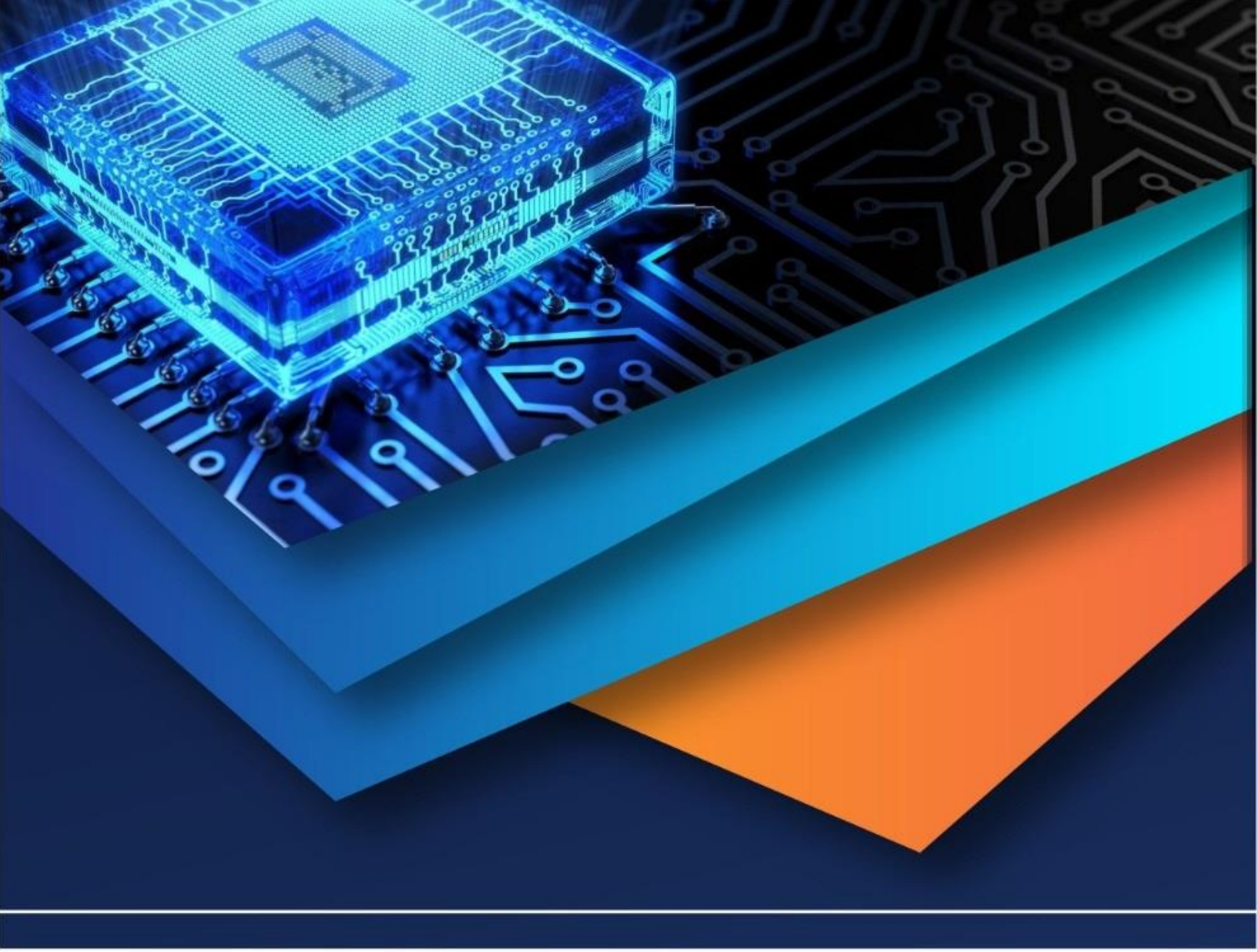
Under extremely limited training conditions, DenseNet-121 emerged as the most effective architecture for the prefix-based image captioning task. The overall ranking observed in this experimental setting is as follows:

DenseNet-121 > VGG-16 > ResNet-18 > ViT-B/16.

From a practical perspective, when computational resources or training time are severely constrained, for example rapid prototyping, educational experiments, or low-budget environments, DenseNet-121 appears to be the most suitable frozen backbone choice. However, it is important to note that with extensive training, larger datasets, and fine-tuning, more advanced architectures, such as vision transformers, are likely to outperform traditional CNN-based models.

REFERENCES

[1] Ding, S., Qu, S., Xi, Y., & Wan, S. (2020). Stimulus-driven and concept-driven analysis for image caption generation. Neurocomputing, 398, 520--530.
[2] Chen, S., Jin, Q., Wang, P., & Wu, Q. (2020). Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9962--9971).
[3] Zhou, Z., Zhang, X., Li, Z., Huang, F., & Xu, J. (2022). Multilevel attention networks and policy reinforcement learning for image caption generation. Big Data, 10(6), 481--492.
[4] Agrawal, V., Dhekane, S., Tuniya, N., & Vyas, V. (2021, July). Image Caption Generator Using Attention Mechanism. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1--6). IEEE.
[5] Zhao, S., Li, L., Peng, H., Yang, Z., & Zhang, J. (2020). Image caption generation via unified retrieval and generation-based method. Applied Sciences, 10(18), 6235.
[6] Liu, X., & Xu, Q. (2020). Adaptive attention-based high-level semantic introduction for image caption. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(4), 1--22.
[7] Mounika, S., & Vijaybabu, P. (2022). Image caption generator using cnn and lstm. South Asian Journal of Engineering and Technology, 12(3), 78--86.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY