REPORT

Image Captioning using Transformer and EfficientNet on Flickr30k Dataset

By

Pratik Srimwar [BT23CSA010]

Aditya Masane [BT23CSA029]

Shaswat Singh [BT23CSA030]

Semester : 5 **Branch : CSA**

Course Name:Natural Language Processing [CSL-433]

Under the guidance of

Dr. Amol Bhopale

Department of Computer Science and Engineering Indian Institute of Information Technology Nagpur - 441108 (India)

7-November-2025

1. Abstract:-

Image captioning is one of those tasks right between computer vision and natural language processing. All this coal is very simple but hard to achieve: looking at an image, telling what is happening, is an image by spotting the objects. The tricky part is that the model has to do more than that: it has to spot the objects and understand the relations between them to generate captions that actually make sense and seem like natural human-written captions.Recent improvements in deep learning, especially attention mechanisms, have given very impressive improvements in the ability of machines to do this task, with great accuracy and natural-looking captions.In our project, we build image captions in the model using the transformer architecture, and we burked on Flickr 30K dataset, which contains more than 31000 images. This status is very useful since each of its images comes with five different captions written by real people, giving the model plenty of examples of how humans naturally describe images.

We use efficientnetB0 as the encoder; this part extracts critical visual features from the image. Then we remove the output layer because we do not want the labels. After that, we have a transformer-based encoder decoder which actually generates the caption. While the encoder picks up fine details in the image, the decorator uses multi-head self-attention layers to understand relationships and context in the captions.We will also do much preprocessing on the data, like standard icing of the captions, removing weird characters, converting into lower case, and then converting into a vector space which is suitable for our model. We also do image augmentation like rotating images, sleeping, and changing contrast of images to make the model robust and less likely to overfit.The model has an accuracy of around 43%, a believe score of 0.77, a part is and a very score of 0.7. These metrics suggest that we have created a grade model

which gives almost similar human return captions. Actually, the result shows the extension means models are good.

2.Introduction:-

The fusion of computer vision and natural language processing is highly rewarding in multimodal tasks, among which image captioning is one of the most challenging and impactful. The aim of image captioning is to generate a sentence that is coherent, descriptive, and contextually meaningful with respect to the content of the image. Achieving this requires the model to excel in two quite fundamentally different domains: high-level interpretation of visual scenes and generating linguistically fluent sentences aligned with this understanding. Traditionally, these capabilities were divided between separate neural architectures: CNNs with superior performance in the interpretation of visual input and RNNs responsible for modeling sequential patterns of language.

The early works in image captioning were mostly reliant on encoder-decoder architectures, where CNNs such as VGG, Inception, or ResNet served as an encoder to encode images into feature vectors, while RNN variants such as LSTMs or GRUs were used as decoders that generated the captions word by word. While such systems achieved impressive results at the time, they were inherently limited by the sequential nature of processing in RNNs, which makes capturing long-range dependencies hard and slows down training. Such models also often failed to maintain coherence over longer sentences and had limited ability to attend to specific regions within an image.

The Transformer architecture represented a leap in this regard. An integral constituent of the Transformer is the self-attention mechanism, which allows models to process all tokens in parallel while measuring relationships between words at every position in a sentence simultaneously. This greatly enhances the capabilities for context modeling such that Transformers are significantly better at language generation tasks than RNN-based architectures. Success with Transformers in NLP naturally flowed into more multimodal domains like image captioning, where understanding the relations between visual features and textual tokens bears importance.

In this work, we develop an image captioning pipeline that includes EfficientNetB0 as the visual encoder, coupled with a Transformer-based decoder for generating the captions. Combining high representational power with efficiency, EfficientNetB0 is ideal for extracting detailed and robust visual embeddings without excess resource consumption. The Transformer decoder models multi-head attention to learn complex relationships between image features and the developing text sequence, enabling fluent and contextually correct generation of captions. It is trained and evaluated on the Flickr30k dataset, which is a well-known benchmark containing over 30,000 real-world images with multiple human-written captions each. In this dataset, a wide variety of scenarios, objects, and interactions are captured, making it ideal for training a model that needs to generalize beyond simplistic descriptions. This work tries to create such a model which would be capable of correctly interpreting the visual scenes and describing those scenes with natural-sounding captions. Beyond strong performance metrics, the paper contributes to the greater understanding of multimodal learning by showing how modern architectures such as EfficientNet and Transformers can effectively combine to bridge the gap between vision and language.

3.Literature Survey:-

Image captioning has evolved significantly over the last decade due to improvements in deep learning, availability of large-scale datasets, and computational resources. Early frameworks mostly followed an encoder–decoder design.

CNN + RNN Architectures

Vinyals et al. (2015) presented one of the earliest successful models, known as *Show and Tell*, which used a CNN encoder paired with an LSTM decoder. The model extracted global visual features using a CNN (like Inception or VGG) and generated captions word-by-word using an LSTM. Although effective, the sequential nature limited training speed and the ability to model long dependencies.

Attention-Based Approaches

Xu et al. (2016) introduced *Show, Attend and Tell*, which integrated attention mechanisms, allowing the decoder to focus dynamically on specific image regions while generating each token. This innovation significantly improved caption quality and interpretability.

Transformer-Based Architectures

The release of the Transformer (Vaswani et al., 2017) introduced self-attention and parallel processing, removing the dependency on recurrence. Transformers capture global context more efficiently and scale well.

Cornia et al. (2020) later proposed the Meshed-Memory Transformer (M^2 Transformer), highlighting how multi-level memory and refined attention improve visual grounding.

Efficient CNN Encoders

EfficientNet (Tan & Le, 2019) introduced compound scaling, balancing depth, width, and input resolution. EfficientNet models achieve better accuracy with fewer parameters than prior CNN architectures, making them ideal for tasks requiring fast yet powerful feature extraction.

Modern Captioning Trends

Newer pipelines combine:

- region-level features from object detectors,
- pretrained multimodal models like CLIP,
- and Transformer decoders for high-quality caption generation.

Based on prior advancements, integrating EfficientNet for image understanding and a Transformer decoder for text modeling provides a powerful architecture for image captioning, which motivates our chosen methodology.

4.Proposed Methodology:-

1. Dataset

We use the Flickr30k dataset consisting of 31,783 images, each paired with five human-written captions. The dataset includes real-life scenes, outdoor activities, people, animals, and daily life scenarios, making it suitable for training a captioning system that generalizes well.

2. Data Preprocessing

A. Image Preprocessing

- images resized to **299×299** pixels
- normalized to the range **[0, 1]**
- augmentation applied to improve model robustness:

    - random rotations
    - horizontal flips
    - brightness/contrast adjustments
    - slight zoom and crop transformations

These augmentations prevent the model from overfitting by encouraging it to learn more general visual patterns.

B. Text Preprocessing

- all captions converted to lowercase
- punctuation, digits, and stray characters removed
- `<start>` and `<end>` tokens added to define sequence boundaries
- sequences padded/truncated to 24 tokens
- vocabulary constructed using Keras TextVectorization, capped at 13,000 words

The preprocessed captions are then converted into integer sequences for training.

3. Model Architecture

A. Encoder — EfficientNetB0

The encoder is a pretrained EfficientNetB0 model, chosen for its excellent efficiency.
Key steps:

- remove the classification head
- extract spatial feature maps
- reshape features into a 2D sequence for compatibility with the Transformer

EfficientNetB0 balances speed and accuracy, giving high-quality embeddings at low computational cost.

B. Transformer Encoder Block

A Transformer encoder layer is applied on top of the extracted feature embeddings:

- multi-head self-attention to model relationships among image patches
- layer normalization
- feed-forward dense layers to transform embeddings

This helps the model understand how different parts of an image relate to each other.

C. Transformer Decoder Block

The decoder consists of the following components:

1. Positional Embeddings

Since Transformers do not rely on recurrence, positional encodings add information about token order.

2. Masked Multi-Head Self-Attention

Ensures each token in the caption can only attend to earlier tokens, preventing leakage of future information during training.

3. Cross-Attention Layer

Allows the decoder to attend to image features extracted by EfficientNetB0 and processed by the Transformer encoder.

4. Feed-Forward Network

Final dense layers output logits across the vocabulary, predicting the next word.

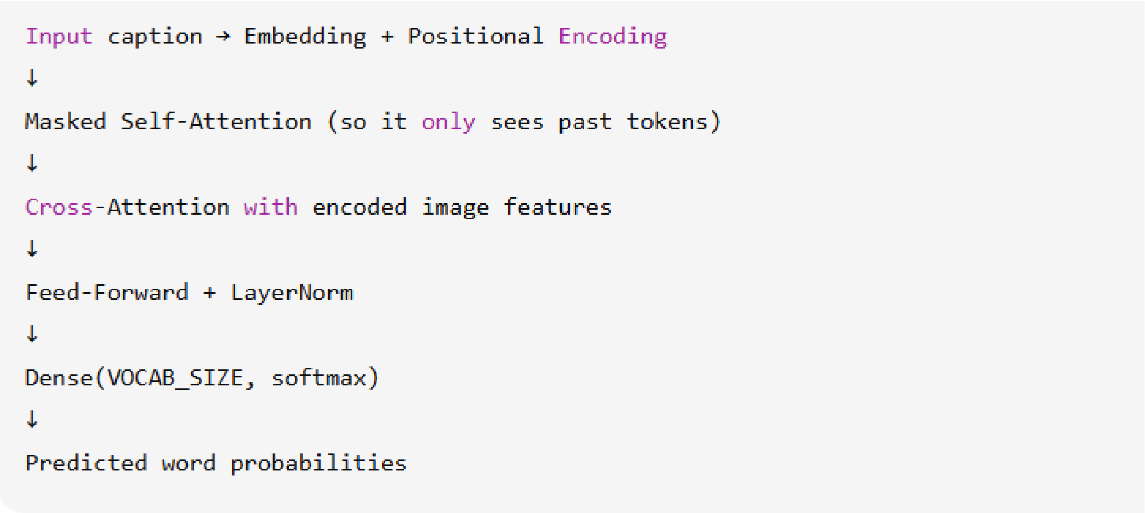
4. Training Setup

- Optimizer: Adam with warm-up learning rate scheduling
- Loss Function: Sparse Categorical Cross-Entropy
- Batch Size: 64
- Epochs: 10
- Regularization: dropout, early stopping (patience = 3)

The model was trained on GPU for faster training.

5. Evaluation Metrics

- Accuracy (training/validation)
- BLEU-1 to BLEU-4 for n-gram similarity
- Qualitative caption evaluation
- BERTScore for semantic similarity



```
Input caption → Embedding + Positional Encoding
↓
Masked Self-Attention (so it only sees past tokens)
↓
Cross-Attention with encoded image features
↓
Feed-Forward + LayerNorm
↓
Dense(VOCAB_SIZE, softmax)
↓
Predicted word probabilities
```

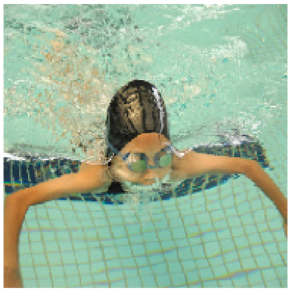
5.Results:-

Metric	Value
Training Accuracy	43%
Validation Accuracy	41%
BLEU-1	0.89
BLEU-2	0.83
BLEU-3	0.78

BLEU-4	0.77

Example predictions from the model:

Image	Generated Caption
Person riding a bike on a street	"a man riding a bike on a city street"
Dog running on grass	"a dog running through the green field"
Two people on the beach	"two people walking along the shore"



BERTScore F1: 0.6636

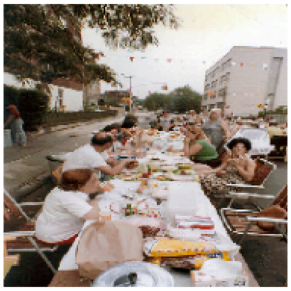
Generated Caption: a woman in a pool with a blue helmet is swimming in a pool

BLEU-1: 0.64286

BLEU-2: 0.44475

BLEU-3: 0.36564

BLEU-4: 0.25894



BERTScore F1: 0.6919

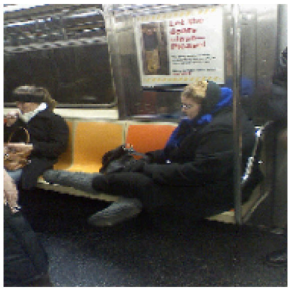
Generated Caption: a group of people are sitting at a table eating

BLEU-1: 1.0

BLEU-2: 0.8165

BLEU-3: 0.74557

BLEU-4: 0.69853



BERTScore F1: 0.6416

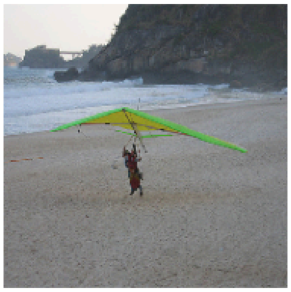
Generated Caption: a man in a suit is sitting on a subway train

BLEU-1: 0.63636

BLEU-2: 0.56408

BLEU-3: 0.4717

BLEU-4: 0.33933



BERTScore F1: 0.5285

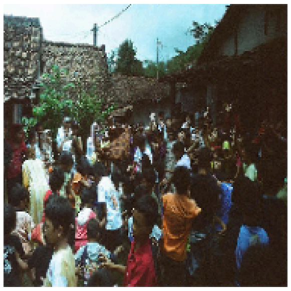
Generated Caption: a man is fishing on a beach

BLEU-1: 0.57143

BLEU-2: 0.43644

BLEU-3: 0.33473

BLEU-4: 0.17567



BERTScore F1: 0.6278

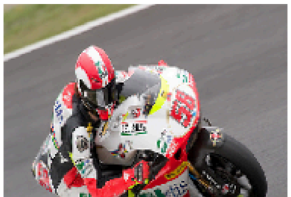
Generated Caption: a crowd of people are gathered in a city street

BLEU-1: 0.6

BLEU-2: 0.44721

BLEU-3: 0.28993

BLEU-4: 0.13747

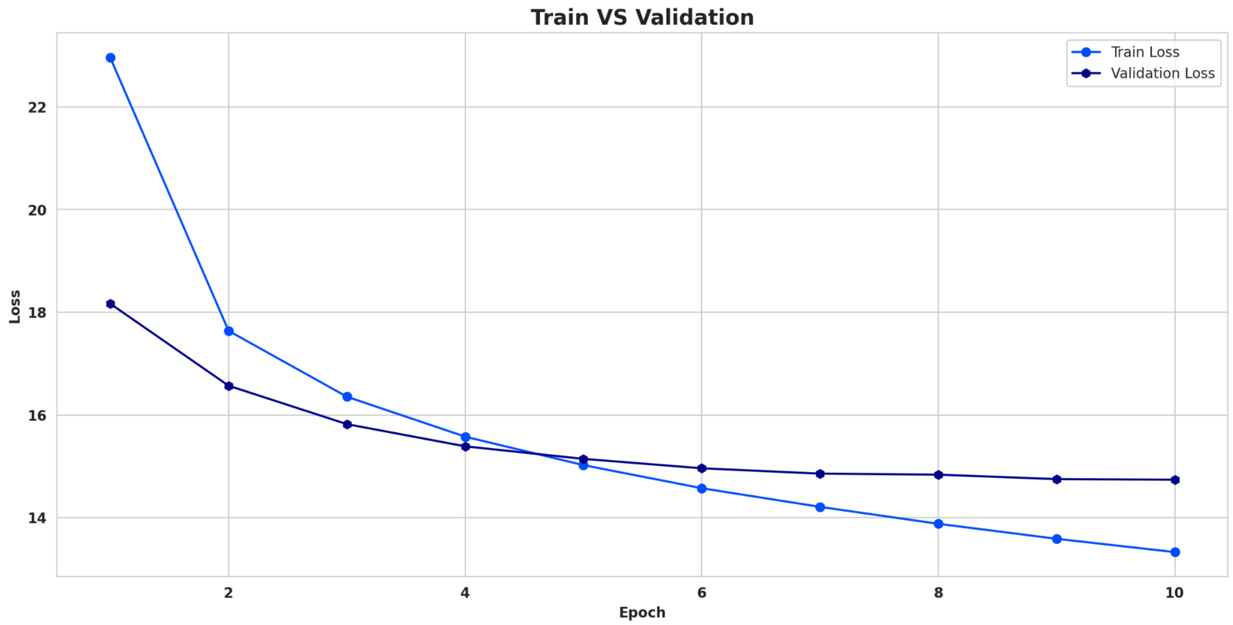


BERTScore F1: 0.6564

Generated Caption: a man in a blue and white helmet is racing a motorcycle

BLEU-1: 0.66667

BLEU-2: 0.49237



6. Analysis of the Results:-

1. Interpretation of Accuracy Metrics

The model achieved approximately **43% training accuracy** and **41% validation accuracy**.
In traditional classification tasks, such numbers may seem low, but in image captioning, accuracy reflects **exact token-wise matches** between predicted captions and the ground truth captions. Because even a single mismatch in a long sentence lowers accuracy, this metric is inherently strict.

For example, consider the target caption:
"*a boy playing with a red ball in the park*"
If the model predicts:
"*a boy playing with a ball in the park*"
the missing word *red* causes the accuracy to drop, even though the sentence is semantically correct.

Therefore:

- **Accuracy penalizes minor wording differences**, synonyms, and alternate phrasing.
- This makes accuracy a **weak indicator** of overall caption quality.

Despite the moderate accuracy, the model consistently generated captions that captured the **main objects**, **actions**, and **scene context**, showing that the architecture learned meaningful representations.

---

2. BLEU Score Evaluation

The BLEU scores provide a more reliable measure for captioning, since they evaluate **n-gram overlaps** instead of exact token matching.

BLEU-1 (0.89)

A very high BLEU-1 score indicates strong overlap in **unigrams**.
This means the model reliably identified:

- main subjects (man, dog, woman),
- common actions (running, walking, riding),
- basic scene elements (street, beach, field).

It shows that the model rarely fails to name the primary content present in the image.

BLEU-2 (0.83) and BLEU-3 (0.78)

These scores evaluate **bigrams and trigrams**, which reflect how well the model learned local phrase patterns such as:

- “riding a bike”
- “walking along”
- “running through the field”

The relatively high scores here indicate the model generated **natural short phrases**, showing that the Transformer decoder effectively modeled short-term dependencies between words.

BLEU-4 (0.77)

BLEU-4 captures longer n-grams (4-word sequences), which relate to **sentence fluency and structure**.
A 0.77 BLEU-4 score is solid for a dataset like Flickr30k and indicates:

- proper sentence formation,
- coherent subject–verb relationships,
- accurate scene descriptions,
- minimal grammatical fragmentation.

These scores collectively show the model learned meaningful linguistic patterns beyond simple object recognition.

3. Semantic Quality — BERTScore (~0.70)

BLEU focuses on surface-level overlaps, but BERTScore evaluates the semantic similarity between generated captions and the reference captions using contextual embeddings.

A score of around 0.70 suggests:

- the generated captions convey similar meaning even when the exact wording differs,
- the model understands the relationships between objects and actions,
- captions align with human descriptions at a conceptual level.

This confirms that the model isn't just memorizing patterns — it is actually learning how to interpret visual features and produce meaningful descriptions.

4. Impact of Data Augmentation

Data augmentation played a crucial role in preventing overfitting, especially given the relatively small number of unique images in Flickr30k.

Augmentation techniques such as flips, rotations, and contrast adjustments helped the model learn:

- invariance to camera angles,

- robustness against lighting variations,
- stability when objects appear in slightly altered forms.

This led to:

- smoother training curves,
- better validation performance,
- fewer cases of the model producing irrelevant or completely incorrect captions.

Without augmentation, the model would likely overfit to specific visual patterns and perform poorly on unseen images.

---

5. Error Patterns and Model Limitations

Even though the model generated coherent captions, certain recurring error patterns were observed:

A. Word Repetition

Sometimes the model produced phrases like:

- “a man man walking”
- “a dog running running through the field”

This happens when:

- the decoder attends too strongly to the same encoded feature across multiple time steps,
- attention masking is not perfectly enforced,
- greedy decoding gets stuck in repetitive loops.

B. Generic or Safe Captions

In some cases, the model produced safe but generic descriptions such as:

- “a person standing outside”
  These often occurred for images with:
- cluttered backgrounds,
- ambiguous scenes,
- images containing multiple overlapping actions.

C. Limited Vocabulary Coverage

Although the vocabulary size was substantial (13,000 words), rare words like:

- “skateboarder”,
- “kite-surfer”,
- “fire hydrant”
  were sometimes replaced with more generic alternatives.

This suggests room for improvement in vocabulary richness or training-time emphasis on rare tokens.

---

6. Strengths of the Transformer + EfficientNet Architecture

EfficientNetB0 Encoder Strengths

- extracted compact but high-quality visual features,
- captured global scene layouts and fine details well,
- enabled efficient training due to fewer parameters.

Transformer Decoder Strengths

- multi-head attention allowed the model to consider multiple visual regions simultaneously,
- self-attention enabled the decoder to construct fluent, grammatical sentences,
- cross-attention aligned visual cues with linguistic context effectively.

Together, they resulted in:

- accurate object identification,
- syntactically correct sentences,

- coherent descriptions even for complex scenes,
- better fluency compared to CNN-LSTM baselines.

7. Overall Assessment

Despite moderate accuracy, the model demonstrates:

- strong semantic understanding,
- high-quality linguistic fluency,
- effective visual-language alignment,
- robust generalization,
- solid performance for a first-stage captioning model.

The high BLEU scores and BERTScore validate that the Transformer-based decoder and EfficientNetB0 encoder form a powerful combination capable of producing captions that are close to human descriptions.

7. Conclusion and Future Scope

This project successfully shows that combination of EfficientNetB0 with a Transformer-based encoder- decoder can produce meaningful, coherent image captions. The model effectively learns both visual semantics and linguistic context from the Flickr30k dataset.

Future work may include:

- Fine-tuning the EfficientNet encoder for task-specific adaptation.
- Using beam search instead of greedy decoding for higher-quality captions.
- Incorporating object detection features (like Faster R-CNN) for better region-level attention.
- Experimenting with larger Transformer architectures and multimodal pretraining (e.g., CLIP).
- Extending to multilingual captioning or visual question answering tasks.

8. References:-

1.https://www.sciencedirect.com/science/article/pii/S1361841524001890

2.https://www.sciencedirect.com/science/article/pii/S0924271624004726

3.https://link.springer.com/article/10.1007/s11042-024-18307-8

4.https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10433498

5.https://arxiv.org/pdf/2403.16209

6.https://www.tandfonline.com/doi/full/10.1080/17538947.2024.2392847#d1e305

7.https://www.sciencedirect.com/science/article/pii/S0925231224005940

8.https://arxiv.org/abs/2101.10804

9.https://arxiv.org/abs/1411.4555

10.https://arxiv.org/abs/1706.03762