CS 772 – FINAL PROJECT EVALUATION

Lavinia Nongbri,23D0383

Prateek Jain, 23M0760

Udhay Brahmi,23M2107

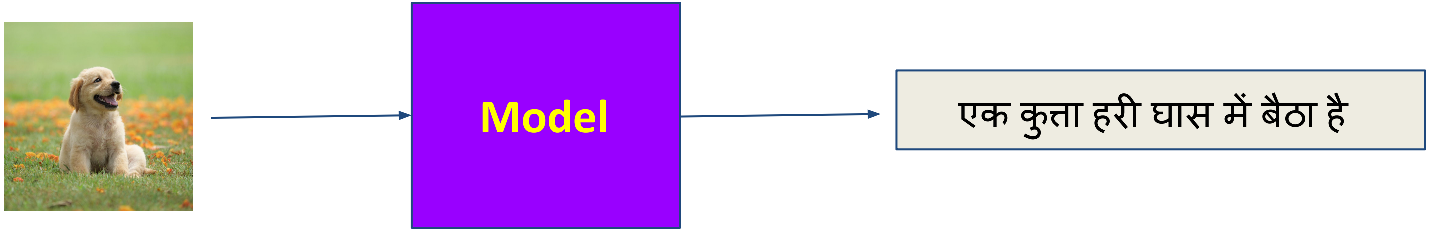
Hasmita Kurre,23D0385

5th may, 2024

Evaluating Image Captioning Methods for Hindi

Input: An Image

Output: Hindi Caption



Problem Statement

- **Objective**: To evaluate and compare the effectiveness of two distinct approaches for generating Hindi image captions using BLEU scores.
- **Methods:**
    - **Direct Captioning in Hindi:** Images are directly captioned in Hindi using a dedicated image captioning model.
    - **Two-Step Captioning via Translation:**
        - Step 1: Images are initially captioned in English using an English image captioning model.
        - Step 2: These English captions are then translated into Hindi using Google Translate.

Motivation for the problem

- **Cultural Relevance:** Effective image captioning in Hindi enhances content accessibility for Hindi-speaking populations, promoting inclusivity in digital media.
- **Technical Challenge:** Developing accurate and context-aware captioning models poses significant computational challenges, especially in languages with fewer resources like Hindi.
- **Research Contribution:** Addresses a gap in current research focused predominantly on English, contributing to the diversification of language technologies in AI.

Literature Survey

- ➢ Rathi, Ankit. "Deep learning approach for image captioning in Hindi language." In 2020 international conference on computer, electrical & communication engineering (ICCECE), pp. 1-8. IEEE, 2020.
- ➢ Papineni, Kishore, 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, pp. 311-318. 2002.

Data Handling (1/2)

Dataset Overview and Preprocessing

- Utilizes the **Hindi-vision-genome dataset**, adapted for Hindi captions. (dataset link)
- Contains **29k images** with hindi and english captions.
- **Dataset format** : <image_id, english caption, hindi caption>
- Preprocessing includes **resizing images** and **tokenizing captions** to prepare data for both direct and translated captioning methods.

Data Handling (2/2)

Dataset Sizes Evaluated

- Evaluated across **different dataset sizes** to assess scalability and robustness: 1k, 5k, 10k, 15k, 20k images and their corresponding hindi/english captions.
- Captions stored in **.txt format** for efficient processing and accessibility during training and evaluation.

Mathematical modelling of the problem

Feature Extraction Equation

$$\mathit{feature} = (W.I + b)$$

Where I represents the input image, W and b are the weights and biases of the model, and $\mathit{feature}$ denotes the feature vector extracted by the Inception v3 model.

Caption Generation Equation

$$C = RNN(\text{feature})$$

Here, C represents the generated caption, and RNN denotes the Recurrent Neural Network (likely LSTM) used to translate the feature vector into a coherent caption.

Mathematical modelling of the problem

Feature Extraction

```
inception = models.inception_v3(pretrained=True)
```

```
self.my_inception = MyInceptionFeatureExtractor(inception)
```

```
features = self.my_inception(images)
```

Caption Generation

```
features = self.encoder(images)
```

```
outputs = self.decoder(features, captions)
```

Methodology/architecture (1/2)

- **Using the Inception v3 Model:**
  - **Purpose**: To extract robust feature vectors from images.
  - **Configuration**: Pre-trained on ImageNet, output feature vectors suitable for caption generation tasks.

- **Encoder Model:**
  - **Implementation**: Custom `Encoder` class that incorporates the Inception v3 model.
  - **Function**: Transforms images into a consistent tensor format for feature extraction, crucial for subsequent decoding into captions.

Methodology/architecture (2/2)

- **Caption decoder:**
  - **Architecture**: Likely includes an RNN or LSTM to generate captions from the encoded image features.
  - **Integration**: Works in tandem with the encoder, converting visual features into textual captions.

```
features = self.encoder(images)
outputs = self.decoder(features, captions)
```

Training Process (1/2)

- **Data Loaders**
  - **Function**: Facilitate the efficient loading, batching, and shuffling of image-caption pairs, crucial for training the neural network.
  - **Implementation**: Customized to handle varying data sizes and ensure consistent input to the model.

Training Process (2/2)

- **Training Loop**
  - Forward pass through the encoder to extract image features.
  - Features fed into the decoder to generate captions.
  - Loss calculation (typically cross-entropy) based on the difference between generated and actual captions.
  - Backpropagation and parameter updates.

Experimental details (1/2)

- **Model Selection**
  - **Primary Model**: Inception v3 for feature extraction due to its proven effectiveness in handling complex image data.
  - **Decoder Model**: Custom LSTM network designed to generate coherent captions based on the features provided by the encoder.
- **Hyperparameters**
  - **Learning Rate**: Initially set to 0.001, with adjustments made based on validation performance.
  - **Batch Size**: 64 images per batch, balancing computational efficiency and training stability.

Experimental details (2/2)

- **Epochs**: Models are trained for up to 12 epochs with early stopping based on validation loss to prevent overfitting.

Metrics

- **BLEU Scores**: Used to quantitatively evaluate the quality of the captions at various n-gram levels, providing a comprehensive measure of linguistic accuracy and fluency.
- **Loss Metric**: Cross-entropy loss is used to measure the difference between the predicted captions and the actual captions, guiding the optimization of the model parameters.

Qualitative Analysis

- **Adequacy:** Measures whether the information in the image is conveyed in the generated caption, regardless if it is fluent or not.

- **Fluency:** Measures whether the generated caption is fluent, regardless of the correct meaning.

- **Score of Adequacy:** Poor, Bad, Moderate, Good, Excellent.

Qualitative Analysis

Generated Caption: एक लड़का एक छोटे से सफेद कुत्ते के साथ खेलता है।

- **Adequacy:** Bad
- **Fluency:** Excellent
- Descriptive phrases like एक छोटे से सफेद कुत्ते
- Misidentification of woman as लड़का and child as कुत्ते

Qualitative Analysis

Generated Caption: एक लाल शर्ट में एक महिला एक सफेद और सफेद कुत्ते के साथ एक सफेद बाड़ के पास एक मैदान

- **Adequacy:** Poor
- **Fluency:** Moderate
- Descriptive phrases like लाल शर्ट, सफेद बाड़
- None of the information in the image is preserved in the captions.

Quantitative Analysis

- **Purpose**: To quantitatively assess the quality of generated captions at various levels of granularity (BLEU-1 to BLEU-4).

- **Methodology**: Compares the machine-generated captions against reference captions to compute similarity scores, providing insights into the model's linguistic accuracy.

Hindi	BLUE-1	BLUE-2	BLUE-3	BLUE-4
1K	0.281254	0.011177	0.000000	0.000000
5K	0.291022	0.015300	0.003316	0.000000
10K	0.301174	0.014558	0.003301	0.000000
15K	0.304047	0.014600	0.000000	0.000000
20K	0.297185	0.145974	0.092793	0.000000

English	BLUE-1	BLUE-2	BLUE-3	BLUE-4
1K	0.341030	0.030419	0.000000	0.000000
5K	0.301900	0.016164	0.003545	0.000000
10K	0.311681	0.032676	0.000000	0.000000
15K	0.311688	0.025796	0.008030	0.001360
20K	0.297185	0.145974	0.092793	0.000000

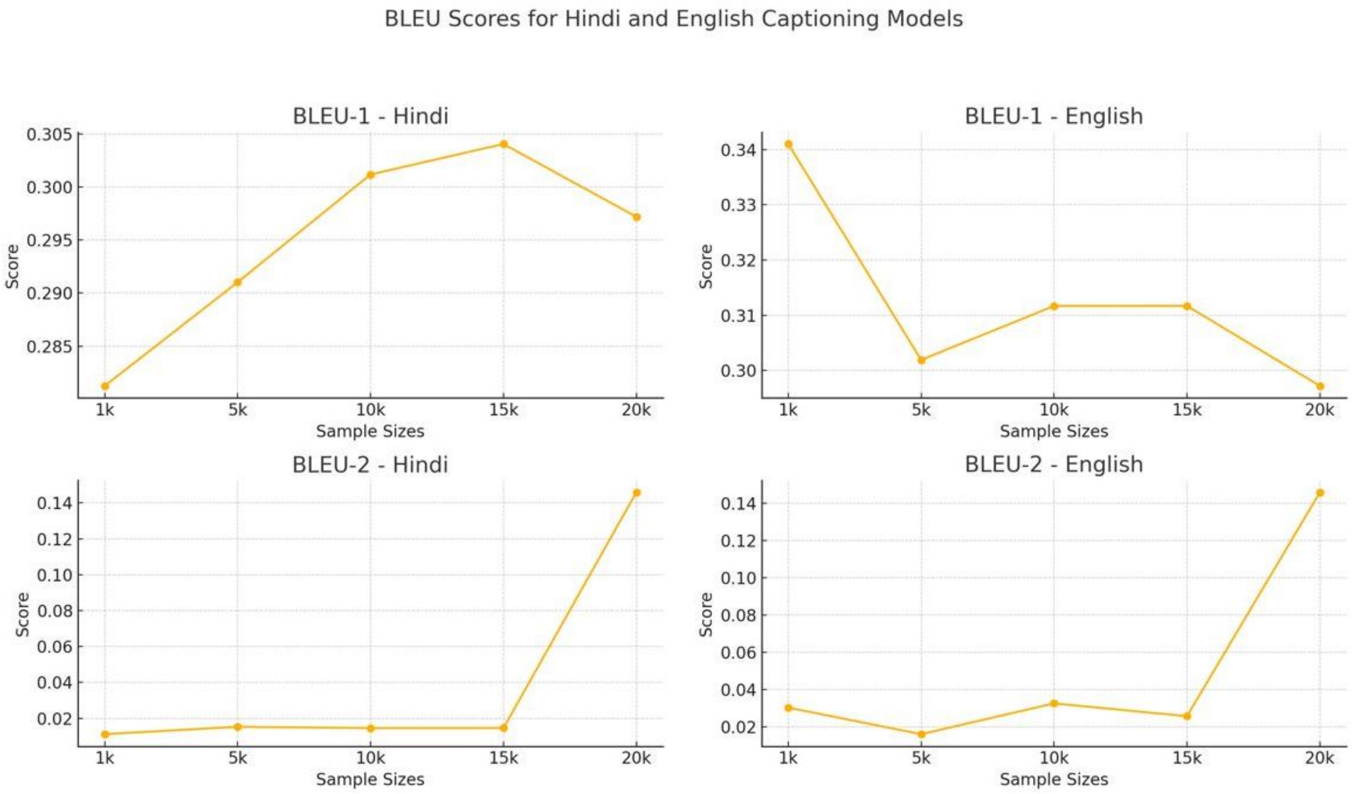
Observation

- **Higher Scores for Single-word Matches (BLEU-1)-**
  - Both models perform best in BLEU-1, which measures the match of single words between the generated captions and references. This suggests that while the models capture common words well, they struggle with more complex linguistic structures.

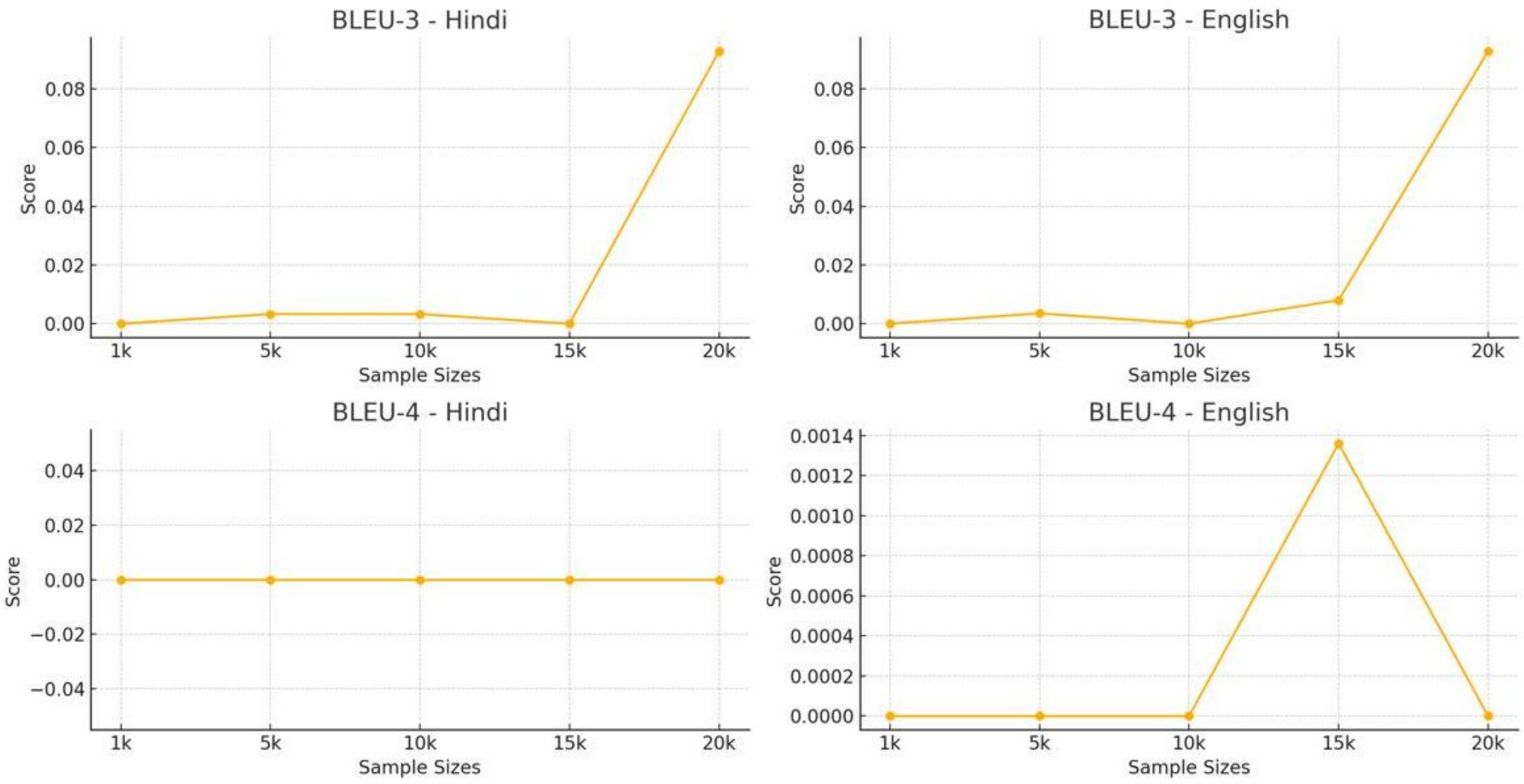
- **Decline in Higher Order n-grams:**
  - BLEU-2, BLEU-3, and BLEU-4 scores, which measure longer matching sequences of words, are significantly lower, indicating challenges in generating coherent longer phrases and sentences.

- **Consistency Across Different Sample Sizes:**
  - As the number of samples increases, there isn't a consistent improvement in BLEU scores, suggesting that simply increasing dataset size doesn't linearly improve performance, especially for higher-order n-grams.

Graphs and Visuals



Graphs and Visuals



Case studies

Generated Caption: एक छोटा कुत्ता एक बड़ी छड़ी के साथ खेलता है।

Generated Caption: एक लाल शर्ट में एक महिला एक सफेद और सफेद कुत्ते के साथ एक सफेद बाड़ के पास एक मैदान

BONUS (Exceeds expectation)

- **Advanced Technological Integration:** The use of a pre-trained Inception v3 model combined with LSTM for generating captions in Hindi, which is less common in computational linguistics, especially for non-Western languages.
- **Innovative Problem-Solving Approach:** Addressing the challenge of direct Hindi captioning alongside a translation-based method, providing a comparative study that enhances understanding of multilingual captioning systems.

Thank You