


Neural rimagar

Image Captioning Model Development and Evaluation

This document presents a comprehensive overview of a project aimed at developing an image captioning model that automatically generates descriptive captions for images. The model integrates computer vision and natural language processing techniques, leveraging datasets, feature extraction methods, and advanced neural network architectures. The project was completed as part of a Natural Language Processing course requirement, submitted on May 2, 2025. The document covers dataset details, preprocessing, model architectures, training procedures, evaluation metrics, and translation of captions into Arabic.

 by **Hossam Mohamed**

A majestic captition with lapers

Dataset and Preprocessing

The project utilizes the Flickr8k dataset, consisting of 8,091 images each paired with five captions. Due to computational constraints, a subset of 1,000 images was selected for training and evaluation. Images are stored in a dedicated directory, and captions are provided in a CSV file with image filenames and corresponding text descriptions.

Captions were preprocessed by converting to lowercase, stripping whitespace, and adding start and end tokens to each caption. A tokenizer was initialized and fitted on the captions to build a vocabulary, with adjustments made to handle special tokens. The vocabulary size and maximum caption length were calculated to guide model input dimensions.

Images were processed using transformations including resizing, normalization, and conversion to tensors. Feature extraction was performed using pretrained convolutional neural networks such as ResNet18 and VGG16, with features saved for efficient training. Data loaders and batching strategies were implemented to facilitate model training.

Model Architectures and Training

Several model architectures were explored for image captioning. Initial models combined ResNet18 feature extraction with LSTM networks to predict the next word in a caption sequence. The architecture included embedding layers, LSTM layers with dropout, and dense output layers with softmax activation. Models were compiled with categorical cross-entropy loss and optimized using Adam.

Advanced models incorporated attention mechanisms, such as additive attention layers, to better align image features with caption generation. Some experiments used Vision Transformer (ViT) models for image feature extraction, projecting features to embedding dimensions compatible with transformer decoders.

Training was conducted with early stopping callbacks to prevent overfitting, using validation splits for monitoring. Batch sizes and epochs were adjusted for efficient training. Model checkpoints saved the best-performing weights. Training progress was monitored via loss metrics, showing steady improvement across epochs.

Caption Generation and Beam Search

Caption generation during inference employed beam search algorithms to explore multiple candidate sequences and select the most probable caption. The beam search incorporated parameters such as beam width, temperature scaling, and top-k filtering to balance exploration and exploitation of word predictions.

Special handling was implemented to avoid common nonsense words and repeated tokens, improving the quality and coherence of generated captions. The generation process stopped upon reaching end tokens or maximum length constraints, ensuring concise and relevant captions.

Example captions were generated for random test images, demonstrating the model's ability to produce descriptive and contextually appropriate text. Captions were displayed alongside their corresponding images for qualitative evaluation.

Evaluation Metrics and Results

The model's performance was quantitatively evaluated using BLEU scores, a standard metric for assessing the quality of generated text against reference captions. BLEU-1 and BLEU-2 scores were computed over a sample of 100 test images, measuring unigram and bigram overlaps respectively.

Despite the model's architecture and training efforts, BLEU scores were low, indicating challenges in generating highly accurate captions. Warnings about zero counts of higher-order n-gram overlaps suggested room for improvement in language modeling and sequence generation.

Qualitative analysis of generated captions alongside original captions and images provided insights into model strengths and limitations. These evaluations informed subsequent improvements in data preprocessing, model design, and training strategies.

Dataset Handling and DataLoader Implementation

To facilitate efficient training, a custom dataset class was implemented using PyTorch, handling image loading, transformation, and caption numericalization. A vocabulary class was built to tokenize captions with frequency thresholds, mapping words to indices and vice versa.

Data loaders were created with batch processing and custom collate functions to pad captions dynamically within batches. This ensured uniform input sizes for model training while preserving data integrity. Batch sizes of 16 were used, with shuffling and multiple worker threads to optimize data throughput.

Initial tests confirmed the correct shapes of image and caption batches, and the total number of loaded images was verified to be 1,000, matching the subset selection. These components formed the foundation for subsequent model training and evaluation.

Transformer-Based Image Captioning Model

A transformer-based image captioning model was developed using a pretrained Vision Transformer (ViT) for image feature extraction. The model projected ViT features to embedding dimensions and employed transformer decoder layers with multi-head attention for caption generation.

The model included embedding layers for captions and a linear output layer mapping to vocabulary size. Training was performed on GPU when available, using cross-entropy loss and Adam optimizer. Over 10 epochs, the model demonstrated decreasing loss, indicating learning progress.

Caption generation functions were adapted to the transformer architecture, producing captions token-by-token with early stopping upon reaching end tokens. Evaluation using BLEU scores was conducted, though scores remained low, highlighting the complexity of the task and potential for further refinement.

Arabic Translation of Captions and Sequence-to-Sequence Model

In addition to English caption generation, the project included a sequence-to-sequence model with attention for translating English captions into Arabic. Character-level tokenization was employed for both input and output texts, with start and end tokens marking sequences.

Data preparation involved one-hot encoding of characters and padding sequences to maximum lengths. The model architecture consisted of LSTM encoder and decoder layers with additive attention, trained using categorical cross-entropy loss and early stopping.

Beam search decoding was implemented for inference, generating Arabic translations of English captions. Sample translations were produced for initial captions, demonstrating the model's capability to perform cross-lingual captioning. The integration of translation enhances the accessibility and usability of the image captioning system.