Visual Question Answering (VQA) Implementation with Attention Mechanism CVPR Proceedings

Rashi Tyagi
University of Illinois Urbana-Champaign
Champaign, IL
rtyagi4@illinois.edu

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

VQA-Project: The system shall interpret an image and answer related questions in natural language. Implementation of "Show, Attend and Tell" uses computer vision and natural language processing combined with attention mechanisms. Pretrained ResNet-50 was used for feature extraction from images, while an LSTM-based decoder with attention was implemented to generate the answers. The system was thus trained and evaluated on the subset of the Flickr8k dataset, achieving accuracy of 90.60% with a BLEU score of 0.5545. This report discusses the details of implementation, results of the project, and inferences.

1. Introduction

The aim of this project is to implement a VQA system that can represent the content of an image and answer questions related to that image in natural language. The VQA task has joined two important fields: computer vision and NLP, aiming to interpret visual and textual data simultaneously. Inspired by the "VQA: Visual Question Answering" paper, the system adapts techniques to process the Flickr8k dataset, which pairs images with natural language captions. VQA remains a challenging task since the model has to integrate visual perception with the understanding of language. Applications include, among others, the accessibility of visually impaired people, the automatic photo analysis, and educational purposes. Key goals of the project are the following:

- Image feature extraction via a pre-trained ResNet-50 model.
- Natural language questions processed through an LSTM-based sequence generator.
- Attention mechanisms implement to focus on relevant regions within an image.
- Evaluating performance with metrics including, but not limited to, accuracy and BLEU score.

The project also borrows ideas from the "Show, Attend and Tell" framework, which introduced attention mechanisms into encoder-decoder architectures for image captioning. Methods have been adapted to answer visual questions, extending the notion of attention to effectively incorporate both visual and linguistic contexts.

1.1. Attention Mechanism Evolution in VQA

Attention mechanisms have turned a new leaf in both computer vision and natural language processing. Initially, the concept came in the context of machine translation, where it allowed for a model to focus on certain parts of an input sequence while generating a corresponding output. This was further extended to vision tasks by the "Show, Attend and Tell" framework, where attention was used to dynamically highlight the relevant regions of an image while generating the caption (Xu et al., 2015). Attention mechanisms were thus integrated into encoder-decoder architectures, significantly enhancing their interpretability and performance.

Attention mechanisms are particularly crucial in VQA to bridge the visual and linguistic modalities. Attention allows the model to answer questions more precisely and contextually by focusing on particular regions of the image that relate to the input question. For instance, Agrawal et al. (2015) used attention mechanisms in the "VQA: Visual Question Answering" paper to enhance the system's ability to point out areas of interest in images when answering complex queries. This adaptation has made the VQA systems more robust and interpretable, thus laying the foundation for the present work.

2. Dataset

Flickr8k was chosen for its relatively manageable size and the presence of descriptive captions for its images. In summary:

- Dataset Size: 8,000 images, each with five captions.
- Image Source: Naturalistic pictures covering nature, cities, human activities, among other subjects.
- Captions: The descriptive nature of the captions provides useful salient features in the pictures and helps to understand them visually and linguistically.
- Preprocessing:
  - Images are resized to 256 x 256 pixels and then normalized.
  - Captions were tokenized into words, and a vocabulary was built with special tokens (,,).
  - A subset of 2,000 image-caption pairs was used for efficient experimentation.

This dataset offers a diverse yet manageable scope for training and testing VQA systems, especially for the evaluation of attention mechanisms.

3. Approach

Comprehensive Methodology

The project followed a systematic approach, emphasizing effective implementation and analysis:

- Framework Selection: The "Show, Attend and Tell" framework was chosen for its success in combining visual and linguistic data through attention mechanisms.
- Model Design: ResNet-50 was utilized for extracting image features, while LSTM processed questions as sequential data. Attention mechanisms bridged these components to focus on relevant regions of images dynamically.
- Hyperparameter Tuning: Hyperparameters such as learning rate (1e-4), batch size (32), and embedding size were optimized through empirical experimentation to balance model complexity and performance.
- Evaluation Metrics: Metrics such as accuracy and BLEU score were selected to assess the system's performance. While accuracy reflects overall correctness, BLEU measures the overlap of generated and reference text using n-grams.

3.1. Data Preprocessing

- Image Processing: Images were resized to 256x256 pixels and then normalized to increase consistency for training.
- Text Tokenization: The captions and questions were tokenized into individual words to enable a mapping to vocabulary indices.

```
for t in range(sequence_length):
    # Compute attention weights
    context, alpha = attention(encoder_features, decoder_hidden)

    # Combine context vector and input embedding
    lstm_input = concat([embedding[captions[:, t]], context])

    # Pass through LSTM
    h, c = lstm(lstm_input, (h, c))

    # Generate output
    outputs[:, t, :] = fc(h)
```

Figure 1. working of the decoder with attention:

3.2. Encoder

The encoder is a ResNet-50 model that is pre-trained on ImageNet. The final layers are replaced to output a feature map of size (batch_size, num_pixels, 2048) that captures both spatial and semantic features of the image.

3.3. Decoder with Attention

This pseudocode in Figure 1 shows how attention dynamically focuses on relevant image regions while generating answers for each time step. The decoder consists of an embedding layer, LSTM network, and attention mechanism:

- Attention Mechanism: This aligns the decoder hidden states with encoder feature maps to compute the context vectors that allow the model to focus on particular regions in an image.
- LSTM Decoder: Processes the concatenated image contexts and embedded question tokens to predict the answer sequentially. LSTM - Long Short-Term Memory networks are good for sequence prediction tasks because they can retain information for a longer period of time.
- Output Layer: This maps LSTM outputs to vocabulary probabilities through a fully connected layer.

3.4. Formulas and Training

Formulas

- Attention Alignment:

$$e_{ij} = W_a \cdot \tanh(W_e \cdot f_i + W_d \cdot h_j)$$

- Attention Weights:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})}$$

- Context Vector:

$$c_j = \sum_i \alpha_{ij} \cdot f_i$$

- BLEU Score Calculation:

$$BLEU = \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log(p_n)\right) \cdot BP$$

  - BP: Brevity Penalty
  - (p_n): Precision of n-grams

- Accuracy:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total predictions}}$$

Training

- Loss Function: Cross-entropy loss minimized token prediction errors.
- Optimizer: The Adam optimizer with a learning rate of 1e-4 was used.
- Epochs: The model was trained for 10 epochs using a batch size of 32.

4. Results

Computational Complexity

Computational demands of the visual question answering tasks necessitated the use of a subset of the Flickr8k dataset instead of larger datasets like MS COCO, which would have required immense computational power. Training deep learning models for computer vision and NLP tasks can be time-consuming, even on NVIDIA T4 GPUs, commonly used in research. This project's training took about 5 hours for 10 epochs on a T4 GPU; at peak, it exceeded 10GB of memory.

Upscaling the present model to much larger sets or using complex architecture would eventually require hardware at least as advanced as NVIDIA A100 GPUs, that offer much greater computational throughput and memory bandwidth for faster training and experimenting with more substantial models like transformers. Future scalability can also be addressed using techniques such as distributed training, mixed-precision computation to optimally use available resources.

Analysis for Table 1:

- The rare tokens "sled" and "twin" had much lower accuracy, indicating the need for handling rare words through methods such as data augmentation or dynamic vocabularies.
- Common tokens such as "bicycle" and "dog" had a high accuracy, indicating frequency bias in model predictions.

Metrics

Token	Predicted Correctly	Total Occurrences	Accuracy (%)
sled	0	5	0.00
twin	1	7	14.29
bicycle	8	10	80.00
ball	12	15	80.00
dog	20	25	80.00

Table 1. Results. Error Analysis Table

- Accuracy: 90.60%
- BLEU Score: 0.5545

Example Outputs

- Image: A man riding a bicycle.
  - Question: What is the person riding?
  - Answer: A bicycle.

5. Discussion and Conclusion

The project was able to prove the effectiveness of attention mechanisms in enhancing VQA performance. The key takeaways are:

- Attention Visualization: Attention weights showed that the model consistently focused on relevant regions of the image for answering questions.
- Evaluation Metrics: High BLEU and accuracy scores showed that the system could generate coherent and accurate responses.
- Challenges: Still, few tricky parts with rare tokens exist, because the infrequent word accuracy is low. High computational complexity may have trade-offs for scalability.

Future Work:

- Experiments with transformer-based architectures like BERT.
- Additional training on synthetic image-question pairs to allow for better generalization.
- Fine-grained evaluations: more additional metrics may be used like ROUGE or CIDEr.

6. Individual Contributions

- Data Preprocessing: Resized images and normalized them; tokenized the text.
- Encoder and Decoder: The architecture was designed and trained; the attention mechanism was included in the design.

- Evaluation: Experiments were performed, results were analyzed-accuracy and BLEU score, and qualitative examples were discussed.

[1–4]

References

[1] A. Agrawal, D. Batra, and D. Parikh. Vqa: Visual question answering, 2015. Retrieved from https://arxiv.org/abs/1505.00468. 4

[2] Illinois Department of Computer Science. Flickr8k dataset, 2023. Accessed via https://forms.illinois.edu/sec/1713398. 4

[3] PyTorch Team. Pytorch documentation, 2023. Available at https://pytorch.org. 4

[4] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. Retrieved from https://arxiv.org/abs/1502.03044. 4