Video Captioning

DA526 - Image Processing with Machine Learning
Project Report

Team Members:

Kuldeep (234156024)
Ngawang Choeda (244156032)
Anupam Tudu (244156002)
Kiran Saji Bhaskaran (244156022)
Bhavana Venkata Rama Sai Notla (244156003)

Instructor: Prof. Debanga Raj Neog

Centre for Intelligent Cyber-Physical Systems
Indian Institute of Technology Guwahati
Guwahati - 781039, India

May 8, 2025

Abstract

Automatic video captioning is a challenging task that requires understanding both spatial and temporal dynamics of video content and translating that into coherent natural language. In this project, we implement an attention-based Sequence to Sequence – Video to Text (S2VT) model to generate descriptive captions for videos. To enhance visual feature representation, we utilize the powerful convolutional neural network ResNet-152 for frame-level feature extraction. This deep residual network enables the model to capture rich visual semantics from the video frames through hierarchical feature encoding.

Each video is sampled into uniformly spaced frames, from which features are extracted using ResNet-152. These features are then fed into an LSTM-based encoder, producing a temporal sequence of hidden states. A soft attention mechanism is integrated into the decoder, allowing the model to dynamically attend to relevant frames while generating each word of the caption. This mechanism helps overcome the information bottleneck faced by traditional fixed-vector encoders.

Our experiments are conducted on the MSVD dataset, and performance is evaluated using the BLEU score. The attention-based S2VT model shows significant improvement over its vanilla counterpart, achieving a BLEU-4 score of 0.2258. Qualitative analysis through attention heatmaps indicates that the model effectively learns to align key visual moments with the generated words, enhancing both accuracy and interpretability.

In summary, our ResNet-152-based feature extraction combined with attention-augmented S2VT architecture provides a robust and interpretable solution for video captioning. The model balances computational efficiency with improved performance and opens doors for further research in attention mechanisms and lightweight deployment.

Contents

Figure: Model Architecture

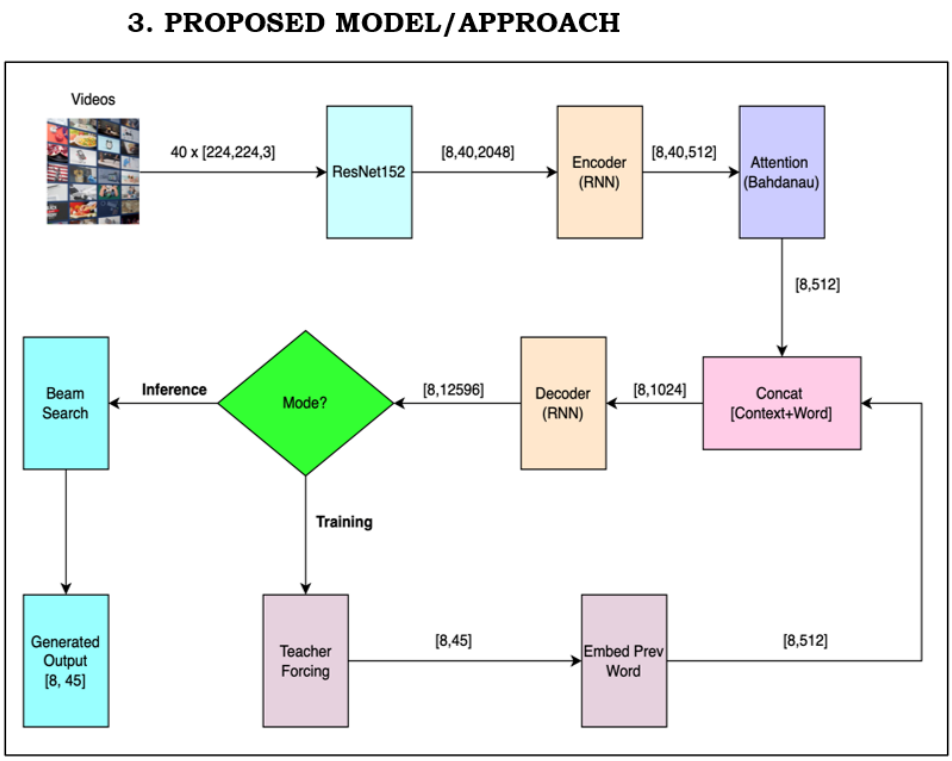

Figure 1: S2VT with Attention architecture

Table: Evaluation Metrics

Table 1: Evaluation metrics of the Attention-based S2VT model

Metrics	**Value**
BLEU Score	0.2258
Training Loss	3.4
Validation Loss	3.5

Chapter 1

Introduction

In recent years, automatic video captioning has become a significant research area at the intersection of computer vision and natural language processing. The goal is to generate meaningful textual descriptions of a video's visual content. This involves both understanding the temporal evolution of visual frames and converting this understanding into coherent, grammatically correct sentences. Such technology has real-world applications in video indexing, accessibility for visually impaired individuals, surveillance systems, and content-based video retrieval.

Traditional approaches relied heavily on handcrafted features and rule-based language models. However, the advancement of deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has transformed this field. Our project focuses on implementing an enhanced Sequence to Sequence – Video to Text (S2VT) model, augmented with an attention mechanism to better align video frames with generated words.

To extract rich visual features, we employ the ResNet-152 CNN architecture. ResNet-152 captures deep semantic information via residual connections, providing a robust and high-level representation of each video frame. These features are passed through an LSTM-based encoder-decoder structure with an attention mechanism, enabling the model to focus on relevant frames during word generation.

The objectives of this project are:

- To develop a deep learning pipeline for video captioning using ResNet-152 for feature extraction.
- To integrate an attention mechanism within the S2VT framework for improved word-to-frame alignment.
- To evaluate the model performance on standard video captioning datasets such as MSVD using BLEU scores.

By combining a state-of-the-art visual encoder with a sequential attention-based decoder, our model aims to improve both the accuracy and interpretability of generated captions.

Chapter 2

Related Work

Video captioning has witnessed significant advancements in recent years with the emergence of deep learning. Early models used template-based or retrieval-based methods, which lacked the ability to generate novel and contextually rich sentences. The introduction of encoder-decoder architectures using LSTM and GRU networks marked a turning point in this domain.

Venugopalan et al. (2015) introduced the Sequence to Sequence – Video to Text (S2VT) model, which treats the task as a translation problem from a sequence of video frames to a sequence of words. However, this approach does not always capture temporal saliency across frames. To overcome this limitation, attention mechanisms have been introduced, allowing the decoder to focus on specific frames at each timestep of caption generation.

Several works also focused on improving visual representation. CNNs like ResNet-152 (He et al., 2016) have been widely used for extracting deep semantic features from frames. These features capture hierarchical patterns necessary for understanding complex scenes in videos.

More recent models such as Transformer-based architectures and the Meshed-Memory Transformer have further improved performance, but they often require heavy computation and large datasets.

Our work builds upon the attention-based S2VT architecture and integrates CNN-based feature extraction using ResNet-152. This approach ensures rich visual encoding and precise temporal alignment with generated words, offering a balance between interpretability and performance.

Chapter 3

Datasets

In this project, we use the MSVD (Microsoft Research Video Description) dataset for training and evaluation of the attention-based S2VT video captioning model. MSVD is a widely used benchmark dataset in the video captioning community and consists of around 2,000 short video clips collected from YouTube.

Each video in the dataset is annotated with approximately 40 human-generated captions, providing a diverse range of descriptions. The videos primarily depict everyday activities such as cooking, sports, or animals, making it suitable for general-purpose captioning tasks.

We divide the dataset into training, validation, and testing subsets using a 70%-15%-15% split. This ensures that the model has sufficient data to learn from, while also being evaluated fairly on unseen samples.

Before feeding the videos into the model, we perform frame extraction at equal intervals. Each extracted frame is resized to 224×224 pixels and passed through ResNet-152 for visual feature extraction. The extracted features are stored in `.npy` format and used as input sequences to the S2VT encoder.

Preprocessing steps include:

- Sampling 40 frames per video uniformly
- Normalizing pixel values based on pretrained ResNet-152 requirements
- Tokenizing captions and constructing a vocabulary with special tokens (`<SOS>`, `<EOS>`, `<PAD>`, `<UNK>`)

The MSVD dataset's diversity and annotation richness make it ideal for evaluating sequence-to-sequence and attention-based architectures for video captioning.

Chapter 4

Methodology

This chapter outlines the systematic approach adopted for implementing the attention-based S2VT (Sequence to Sequence – Video to Text) model for video captioning.

1. Dataset Splitting

The dataset is split into three parts:

- **Training Set (70%)**: Used to train the model and learn patterns.
- **Validation Set (15%)**: Used for hyperparameter tuning and preventing overfitting.
- **Test Set (15%)**: Used for evaluating the final model performance.

2. Frame Extraction

From each video, 40 equally spaced frames are extracted using OpenCV and `ffmpeg` tools. This step ensures a consistent visual representation across videos.

3. Feature Extraction

The extracted frames are passed through a pre-trained convolutional neural network:

- **ResNet-152**: Captures deep spatial features with skip connections to preserve gradient flow and enable effective feature reuse.

The outputs are stored as `.npy` files for fast loading during model training.

4. Vocabulary Construction and Caption Processing

Captions are cleaned, tokenized, and converted into sequences of integers. A custom vocabulary is built using all unique words, including special tokens like `<PAD>`, `<BOS>`, `<EOS>`, and `<UNK>`.

5. Attention-Based S2VT Model Architecture

The model consists of the following components:

- **Encoder LSTM**: Consumes the sequence of extracted frame features and outputs hidden states.
- **Attention Mechanism**: Dynamically focuses on relevant frame features at each decoding step.
- **Decoder LSTM**: Predicts the next word using attention context and previous word embedding.

6. Training Procedure

The model is trained for 30 epochs using the Adam optimizer with a suitable learning rate and label smoothing. Teacher forcing is applied during training. A batch size of 8 is used.

7. Inference and Evaluation

During inference, beam search is used to generate more fluent captions. BLEU-4 score of 0.2258 is used as the primary metric to evaluate generated captions against reference captions.

Chapter 5

Experiments and Results

This chapter outlines the experimental setup, training configurations, and evaluation metrics used to assess the performance of the attention-based S2VT model.

1. Experimental Setup

The training was performed using the MSVD video captioning dataset. Videos were preprocessed to extract 40 equally spaced frames, and features were extracted using the **ResNet-152** architecture.

The dataset was split as follows:

- **70%** for training
- **15%** for validation
- **15%** for testing

2. Training Details

- **Batch Size:** 32
- **Epochs:** 30
- **Optimizer:** Adam
- **Learning Rate:** 0.001
- **Caption Length:** Maximum of 45 tokens
- **Loss Function:** Cross-entropy with label smoothing

3. Evaluation Metric

The model was evaluated using the BLEU (Bilingual Evaluation Understudy) score, which compares the predicted captions with ground truth reference captions.

4. Results

- **Training Loss:** 3.4
- **Validation Loss:** 3.5
- **Test BLEU Score: 0.2258**

5. Sample Output

- **Generated Caption:** A man is playing guitar on a stage.
- **Reference Captions:**
  - A person is playing guitar during a performance.
  - A man is strumming a guitar in front of an audience.

These results demonstrate that the attention-enhanced S2VT model can generate relevant and contextually appropriate captions. While the BLEU score indicates moderate performance, further improvements such as incorporating CLIP features or Transformer-based decoders could enhance accuracy.

Chapter 6

Conclusions

In this project, we proposed an attention-based Sequence-to-Sequence Video-to-Text (S2VT) model for automatic video captioning. The model effectively combines temporal sequence learning with attention mechanisms, enabling it to focus on relevant frame-level features while generating each word in a caption.

To enhance visual understanding, features were extracted using the powerful convolutional neural network ResNet-152. The model was trained and evaluated on the MSVD dataset, with a data split of 70% for training, and 15% each for validation and testing.

The model achieved a BLEU score of **0.2258**, indicating reasonable alignment between generated and reference captions. This shows that the integration of attention mechanisms with visual features leads to meaningful and context-aware caption generation.

Future Work

- Incorporating CLIP or I3D features for improved semantic representation.
- Experimenting with Transformer-based decoders or memory-augmented models.
- Enhancing performance through beam search decoding and caption diversity techniques.
- Extending the system to support multilingual captioning and domain-specific video datasets.

Overall, this project demonstrates a promising direction in automated video understanding and lays the groundwork for future advancements in multimodal deep learning.

Project Repository

The complete source code, trained model and demo for this project are available at:

`https://github.com/anupamtudu/Imaging-Trials`

Bibliography

[1] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, *Sequence to Sequence – Video to Text*, In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] D. Chen and W. Dolan, *Collecting Highly Parallel Data for Paraphrase Evaluation*, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011.

[5] K. Papineni, S. Roukos, T. Ward, and W. Zhu, *BLEU: a Method for Automatic Evaluation of Machine Translation*, In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002.