Dense video captioning based on local attention

0) ABSTRACT

Problem Addressed: The paper focuses on dense video captioning, which involves locating multiple events in untrimmed videos and generating captions for each event.

Challenges in Prior Methods: Previous approaches struggled with establishing the relationship between video frames and captions, resulting in low-quality generated captions.

Proposed Model - DVCL: The Dense Video Captioning Model Based on Local Attention (DVCL) is introduced to address these challenges.

Feature Extraction: DVCL uses a 2D temporal differential CNN to extract video features effectively.

Global Feature Dependence: It employs a deformable transformer to encode features, establishing global dependencies within the input sequence.

Improved Event Proposals: DIoU and TIoU are integrated into the event proposal match algorithm, enhancing the accuracy of event proposals.

Caption Generation: The model uses an LSTM with local attention to generate captions, ensuring that each word corresponds to the relevant frame.

Experimental Results: Extensive experiments on the ActivityNet Captions dataset show that DVCL outperforms other baseline methods, achieving significant improvements in BLEU4, METEOR, and CIDEr scores.

1) INTRODUCTION

Challenging Task: Video captioning aims to generate meaningful sentences describing the content of a video. Dense video captioning deals with untrimmed, multi-event videos, requiring a sentence for each event.

Two-Stage Process: Dense video captioning involves event proposal and caption generation. Event proposal identifies event segments, and caption generation translates them into natural language.

Challenges: Existing methods often struggle to establish relationships between frames and captions, leading to low caption quality. Recurrent Neural Networks (RNNs) have limitations with long videos, and LSTM-based encoders are inefficient in processing fine-grained motion.

Proposed Solution - DVCL: The Dense Video Captioning Model Based on Local Attention (DVCL) uses a CNN for initial feature extraction from video frames. It introduces an attention mechanism to capture temporal dependencies and a local attention mechanism to improve model speed and caption accuracy.

Key Contributions:

1. Introduction of a local attention-based LSTM to dynamically focus on frame features corresponding to each generated word, improving caption accuracy.
2. Inclusion of DIoU and TIoU in the event proposal match algorithm and evaluation algorithm during training to obtain more accurate event proposals and enhance caption quality.
3. Extensive experiments on ActivityNet and YouCook2 datasets, demonstrating that DVCL outperforms existing baselines and achieves state-of-the-art performance.

2) RELATED WORK

Video Captioning Categories:

- Video captioning methods are categorized into template-based and sequence-based approaches.
- Template-based approaches use predetermined templates for captions, limiting linguistic diversity.
- Sequence-based approaches, employing CNNs and RNNs, generate more flexible sentences.
- Most recent video captioning approaches follow the sequence-based model, leveraging deep learning techniques.

Handling Multiple Events:

- Traditional video captioning focuses on a single event in videos, while most videos contain multiple events.
- Dense video captioning was introduced to address this, considering both short and long events.
- Contextual relationships between adjacent events are used to generate captions.

Challenges in Video Captioning:

- Traditional methods consider event-level contextual relationships but neglect feature relationships between video frames.
- Recent approaches involve event, scene, and frame-level features but miss contextual relationships of graphical text.
- Vision Transformer and end-to-end models have been explored but face challenges of manual components, redundancy, and readability issues.

The Role of Attention Mechanism:

- Attention mechanisms can establish feature relationships between frames and captions.
- The proposed model, Dense Video Captioning Model Based on Local Attention, employs local attention to focus dynamically on input frame features during word generation, resulting in captions with higher accuracy.

3) METHOD

3.1) Future Extraction

Temporal Differential Network:

- Dense video captioning can be computationally intensive due to coding spatiotemporal information of all frames.
- A sparse sampling algorithm is employed to select a limited number of frames distributed uniformly in the video.
- A 2D convolution-based temporal differential encoder captures short-term motion information around moving objects.
- This encoder extracts visual and motion features from the selected frames.
- A network structure includes average pooling, 2D convolution, and max pooling to create feature maps.
- Feature maps are merged with the sampled frames to facilitate long-distance dependence modeling.

Frame Sampling and Feature Fusion:
- Original videos are divided into equal segments, and a frame is randomly selected from each segment.
- Random sampling increases training diversity and allows the network to learn different instances of the same event.
- The selected frames are processed by a 2D convolutional neural network to obtain feature representations.
- Each frame's feature representation includes visual information (Ft) and local motion information (H(xt)).
- Visual information from each frame (Ft) is combined with the local motion information (H(xt)) in a feature stack.
- The feature stack (H(xt)) is generated by extracting features from n frames around the sampled frame and stacking them.
- The final sampled feature sequence is used as the input for the transformer network.

3.2) Feature Encoding
Multi-Head Self-Attention for Feature Encoding:**
- To capture global feature relations in the video, a transformer network with multi-head self-attention is adopted.
- The multi-head self-attention mechanism calculates attention weights for each sampled image frame.
- Attention is calculated by creating queries (q) and keys (k) for each frame and applying a learnable weight matrix (Wm).
- The self-attention weights (Amqk) are determined using the attention formula and normalized.
- However, calculating global self-attention for all frames is computationally intensive and leads to parameter redundancy.
Multi-Head Deformable Transformer Encoder:
- To address computational and redundancy issues, a multi-head deformable transformer encoder is introduced.
- It samples a set of key frames around the current frame, allowing for efficient computation of attention weights.
- The key vectors (k) are assigned to the query vectors (q) for elements in the sequence.
- The attention weights are computed by fully connected layers, making use of sampling offset and attention weight parameters (Δpmqk and Amqk).
- This approach differs from prior methods that used dot products for attention weights.

3.3) Event Proposal
Challenges in Event Proposal Matching:
- Event proposal tasks encounter challenges with blurred event boundaries and sparse labeling.
- Strict one-to-one matching of predicted and real event boundaries is problematic, as multiple candidate intervals can have similar quality.
- Different candidate intervals for the same event prediction have similar accuracy, making it hard to distinguish superiority.
Introducing DIoU and TIoU:
- To address these challenges, Distance Intersect-over-Union (DIoU) and Tightness-aware Intersect-over-Union (TIoU) are introduced in the proposal matching algorithm and proposal evaluation algorithm during event proposal training.
- DIoU focuses on the location and completeness of the event segments, improving event proposal accuracy.
- TIoU ensures event proposals' precision by focusing on completeness and avoiding the use of a threshold that determines the final result.

Proposal Matching Algorithm:
- Event proposals are matched with real events, and interval error is calculated as the loss for optimization.
- Matching function combines Intersection over Union (IoU) and L1 distance to focus on both event overlap and boundary matching.
- DIoU is incorporated into the IoU function to enable the optimization of the distance between predicted and real event intervals, even when they don't overlap.

Proposal Evaluation Algorithm:
- To optimize event proposals, the proposal evaluation algorithm uses TIoU to determine the precision of predicted event intervals more accurately.
- TIoU ensures completeness by focusing on each part of the predicted event intervals and applying a penalty if other events or background noise are included.
- Event proposal evaluation algorithm includes boundary loss (Lbou) to measure deviations in the starting and ending frames and event loss (Lpre) to measure the accuracy and completeness of predicted event intervals.

3.4) Caption Generator

Generating Event Descriptions:
- The predicted events are represented as $\Psi = \{\varphi(\hat{t}s,n, \hat{t}e,n)\}Nn=1$, with N being the number of events in each video.
- Feature sequences {xt,n} for sampled image frames within the event are used for text generation.
- The conditional probability of generating each word in the description text is calculated using an attention-based recurrent neural network.

Conditional Probability Parameterization:
- Conditional probabilities are parameterized, making them expressible as soft-max functions.
- The hidden layer state (hj) of a recurrent neural network (RNN) is computed based on the previous hidden layer output and the current vector.
- The context relationship vector (ct) is introduced and combined with the hidden state to create a new hidden state that carries the attention mechanism.

Attention Mechanism:
- An attention mechanism is added to the caption decoder to focus on hidden state information for generating more readable and accurate descriptions.
- The context relationship vector (ct) dynamically changes according to the relevance of the current prediction sequence with the input sequence in the encoder, computed by assigning different weights to the hidden vectors of the input sequence at each time step.

Network Framework:
- The network framework of the attention mechanism is illustrated, showing how attention is focused on the relevant frames in the sequence for more accurate description generation.

Attention Window Calculation:
- In generating each target word, the center (pt) of current attention is calculated, allowing the output word to pay attention to the position of the input sequence associated with itself.
- The position matrix (Wp), penalty term (vp), and sigmoid activation (sigmoid) are used to determine the attention window.

Calculating Attention Weights:
- The attention weights are obtained by aligning the hidden layer vectors of the input and output sequences using an align function, constrained by a Gaussian distribution.

Generating Output Words:
- The LSTM network takes input from contextual features, attention weights, and previously generated words.
- The output words at the current position are calculated using a fully connected network and SoftMax activation function.

4) EXPERIMENTS

4.1) Dataset &Evaluation Metrics

ActivityNet Captions Dataset:

- This dataset comprises 20,000 videos, each with corresponding captions.
- The average video length is 120 seconds, containing an average of 3.65 events per video.
- The number of events in videos follows a normal distribution.
- Captions have an average of 13.48 words per sentence, also conforming to a normal distribution.
- Approximately 10% of the captions have partial overlap with the corresponding events.

YouCook2 Dataset:

- Contains 2,000 videos depicting 89 cooking scenarios.
- Cooking steps are labeled with start and end positions and described using English imperatives.
- Videos are from YouTube and are shot from a third-person perspective.
- This dataset is used to verify video captioning accuracy in single-event scenarios.

Dataset Splits:

- For ActivityNet Captions: Training/Validation/Test splits are 10,000/4,900/5,000.
- For YouCook2: Training/Validation/Test splits are 1,300/450/210.

Evaluation Metrics:

- The evaluation metrics used for assessing the model's performance are BLEU, METEOR, and CIDEr.
- These metrics are commonly employed in the field of text translation and are used to measure the quality of the generated captions.

4.2) Baselines

DCE (Dynamic Context Event)

- Utilizes C3D as the backbone network to extract video feature information.
- Employs a dynamic step size for event extraction.
- Incorporates an attention mechanism within the LSTM network to establish contextual dependencies between different events, enhancing caption accuracy.

MT (Masked Transformer)

- Employs a transformer network for end-to-end training in video captioning.
- Demonstrates the ability of the self-attention mechanism to establish cross-modal feature relationships between text and images.

DVC (Description-Proposal Video Captioning)

- Predicts event and description scores to filter events with higher complexity as candidates.
- Improves sentence coherence through reinforcement learning.

Efficient

- Reformulates caption generation as a classification problem.
- Given an input sequence and context vector, the model selects the target caption from a series of candidates.
- Enhances caption accuracy through unsupervised learning.

ECHR (Event-Centered Hierarchical Representation)

- Uses an event-centered hierarchical feature representation to extract event-level, scene-level, and frame-level features as input for caption generation.
- Generates sentences by combining feature information at different levels.

PDVC (Parallel Decoding Video Captioning)

- Employs an end-to-end network structure with parallel decoding to generate both candidate events and captions.
- Addresses the issue where the back-propagation of the caption generation model cannot directly guide the optimization of the event proposal model.

4.3) Experiment Setup

Image frame sampling:

- All videos are uniformly cropped to 224 × 224 pixels.
- Each image block is divided into 16 × 16 pixels.
- Frame rate: 1⁄32.

Transformer network:

- Eight self-attention heads.
- 512-dimensional hidden layers.
- 2048-dimensional feedforward neural networks.
- Dropout rate: 0.5.
- Activation function: ReLU.

Caption generation:

- LSTM network structure.
- 512-dimensional hidden layers.
- Local attention mechanism with a window size of 5.
- Final word vector output through a 1-layer, 512-dimension fully connected network with softmax activation function.

4.4) Result Analysis

Comparative Analysis with Baselines:

- DVCL outperforms six baselines significantly on the ActivityNet Captions dataset.
- Achieved optimal results in BLEU4, METEOR, and CIDEr with notable improvements over the best baseline.
- Baselines like DCE, MT, DVC, Efficient, ECHR, and PDVC were compared, highlighting DVCL's strengths.

Analysis of Baselines:

- Detailed analysis of baseline models, their strengths, and weaknesses in handling dense video captioning.
- DVCL compared favorably due to its effective utilization of contextual information and attention mechanisms.

Single-Event Scenario Comparison:

- Demonstrated DVCL's accuracy in generating descriptive text for single-event scenarios.
- Proved its capability to accurately describe videos containing only a single event.

Ablation Studies:

- Analyzed the impact of different components in DVCL on overall model performance and text quality.
- Highlighted the crucial role of the local attention mechanism in caption generation.

Performance Improvements in DVCL:

- Identified key reasons for DVCL's superior performance, including efficient feature extraction, global feature dependencies, and effective attention mechanisms.

5) CONCLUSION

Model Introduction: The paper introduces a novel dense video captioning model called DVCL.
Objective: The model aims to address the issues of low accuracy and poor readability in generated video captions.
Feature Extraction Module: DVCL employs a 2D temporal differential network to effectively extract video features.
Global Feature Dependency: It constructs the global feature dependency of the input sequence using a deformable transformer.
Event Proposal Module: The model uses TIoU and DIoU to generate more accurate event proposals.
Caption Generator Module: DVCL employs an LSTM-based local attention mechanism that dynamically focuses on input frame features, improving caption accuracy.
Experimental Results: The proposed method is evaluated on the ActivityNet Captions and YouCook2 datasets, demonstrating significant improvements in caption quality and accuracy.