VIDEO CAPTIONING WITH ATTENTION-BASED LSTM AND SEMANTIC CONSISTENCY

Authors	Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu and Heng Tao Shen, Senior Member, IEEE
Published by	IEEE
Published in	IEEE Transactions on Multimedia (Volume: 19, Issue: 9, September 2017)
Date of Publication	19 July 2017
Cited in	451(in papers) & 1(in patent)
INSPEC Accession Number	17099707
DOI	10.1109/TMM.2017.2729019

Abstract:

- Recent advancements in LSTM for image captioning inspire exploration for video description using natural language sentences.
- Existing methods compress video frames into a static representation without incorporating attention mechanisms to select salient features.
- This study introduces a novel framework called aLSTMs, integrating attention-based LSTM with semantic consistency for video-to-sentence transformation.
- aLSTMs utilize attention mechanisms with LSTM to capture salient video structures and explore correlations between word embeddings and visual content for rich semantic sentence generation.
- Attention mechanism involves dynamic weighted sum of local 2D CNN representations.

- LSTM decoder combines visual features and word embeddings to generate significant words.
- Multimodal embedding maps visual and sentence features into a joint space to ensure semantic consistency between sentence description and video visual content.
- Experiments on benchmark datasets demonstrate that aLSTMs achieve competitive or superior results compared to state-of-the-art baselines for video captioning in BLEU and METEOR metrics, even with a single feature.

I. **Introduction**:

- Integration of visual content with language learning for image and video description is a challenging task with numerous applications.
- Deep Convolutional Neural Networks (CNN) have significantly advanced image description generation, but video captioning is more complex due to diverse elements.
- Long Short-Term Memory (LSTM) has been successful in language translation and image captioning, inspiring its extension for generating video sentences with rich semantic content.
- Attention networks are essential in deep learning, providing variable-length memory and enabling soft-selection over source elements, enhancing results in various tasks.
- The proposed aLSTMs framework is an attention-based LSTM model with semantic consistency, aiming to capture salient temporal structures of videos and bridge the semantic gap between videos and sentences.
- Inception-v3 neural network and LSTM visual encoder are used to extract spatial and temporal features, respectively.
- Attention mechanism uses dynamic weighted sum of local spatial 2D CNN feature vectors as input for the LSTM decoder.
- Multi-word embedding and cross-view methodology are integrated to project generated words and visual features into a common space, ensuring consistency.
- Key aspects of the approach include attention mechanism, cross-view model for consistency, and achieving competitive results in video captioning.

- The paper covers related work, details of the proposed approach, experimental settings, and results.

II. Relevant Work:

A. Image/Video Recognition:

- Recognition of image and video is a fundamental and challenging problem in computer vision.
- Supervised convolutional models have made dramatic progress in image-based action recognition.
- Pre-trained CNN models have advanced image feature learning, but directly applying them to process videos is challenging due to the lack of dynamic information.
- Approaches like deep 3D CNN and LSTM are used to learn spatio-temporal features for video analysis tasks.
- LSTM helps model sequence data and learn patterns with a wider range of temporal dependencies, aiding spatio-temporal information learning from videos.

B. Image/Video Captioning:

- Bridging the gap between video/image understanding and natural language processing is a significant research focus, aiming to generate sentences describing image/video content.
- Various methods leverage RNN and LSTM to automatically describe images/videos with correct and novel natural language sentences.
- Multi-modal learning and machine translation principles inspire joint multimodal embedding approaches to align image and text features into a common space.
- Several approaches focus on video captioning, leveraging CNN and LSTM to generate captions by considering both temporal and spatial information in videos.
- Attention networks have shown success in embedding categorical inference within deep neural networks, particularly useful in machine translation, visual question answering, and video/image captioning.

- Attention-based models help identify the importance or weights of source words for translation, align relevant visual content with the next word in a sentence, and select the most relevant temporal segments for generating the next word.

III. <u>Proposed Approach</u>:

A. Terms and Notations:

- Video V described by a textual sentence D consisting of Nd words.
- Visual features X and textual features D0 are used to represent the video and sentence.
- Xt extracted using deep neural networks.
- M and L denote the dimension of visual and textual features, respectively.

B. Attention-based Long Short-Term Memory Decoder:

- LSTM is applied as the basic component for the aLSTMs approach.
- LSTM helps model sequence data and learn patterns with a wider range of temporal dependencies.
- Attention mechanism is integrated into LSTM to align temporal information in videos with language data.
- Attention weights are computed to determine the relevance of each feature in the input video.

C. Loss 1: Translation from Videos to Words:

- Defines the likelihood of generating each word in a training sentence given the context vector.
- Defines the cost of generating each word based on negative logarithm of the likelihood.
- Aims to ensure coherence and smoothness in sentence generation.

D. Loss 2: Bridging the Semantic Gap with Semantic Cross-view Correlation:

- Performs one-layer LSTM process over video visual features to generate a single visual feature with spatial and temporal information.
- Performs mean pooling over embedding vectors to construct a single sentence feature for each description.

- Maps visual and sentence features into a common high-level abstract space through linear projection.
- Minimizes cross-correlation to ensure semantic consistency between generated words and video visual context.

E. Attention-based LSTM with Semantic Cross-view Correlation:

- Formulates the video captioning problem by minimizing an energy loss function that balances translation and semantic consistency.
- Utilizes BeamSearch to generate a coherent and precise sentence describing the video.

IV. <u>Experiments</u>:

- The paper describes an experiment in video captioning and provides detailed information on datasets, implementation details, and results. The primary focus of the experiment is to evaluate the performance of an algorithm for generating video captions
- They used two publicly available datasets, MSVD and MSR-VTT, which are commonly used for video captioning tasks. They preprocessed the data, used a Convolutional Neural Network (CNN) to extract visual features, and employed LSTM units for generating captions. The algorithm also introduces a semantic cross-view correlation loss to improve the quality of the generated captions.
- They conducted experiments to study the influence of parameters and compare their results with state-of-the-art methods. The results indicate that their approach outperforms several baseline methods in terms of various evaluation metrics, including BLEU, METEOR, and CIDEr. They also conducted sub-experiments to analyze the effect of different factors, such as the choice of visual encoder and the semantic cross-view correlation.
- The findings show that their approach, which uses an Inception-v3 visual encoder and incorporates semantic cross-view correlation, performed better than other methods. This demonstrates the importance of both spatial and temporal information in video captioning tasks.

- The paper provides specific results and examples to illustrate the performance of their algorithm, highlighting instances where it correctly generated captions, incorrectly identified objects, or used incorrect verbs.
- In summary, the paper presents an algorithm for video captioning that outperforms several state-of-the-art methods, emphasizing the importance of capturing both spatial and temporal information in video analysis.

A. *On MSVD dataset:* (aLSTMs)

- BLEU@1: 81.8%
- BLEU@2: 70.8%
- BLEU@3: 61.1%
- BLEU@4: 50.8%
- METEOR: 33.3%

B. *On MSR-VTT dataset:* (aLSTMs)

- BLEU@4: 38.0%
- METEOR: 26.1%

V. Conclusion:

- In this paper, we introduced a novel framework called aLSTMs, designed to optimize relevance loss and semantic cross-view loss simultaneously.
- By applying this framework to two widely used video description datasets, our experiments showcased the effectiveness of our approach.
- In fact, it achieved comparable or superior performance when compared to current state-of-the-art models.
- Looking ahead, we intend to tailor our model to cater to domain-specific datasets, such as movies, to further improve its capabilities and performance.
- Numeric Metrics (Bleu@4 and METEOR) for the comparison of state-of-the-art algorithms with our proposed method on the MSR-VTT dataset:

    - MP-LSTM (GoogleNet): Bleu@4 = 34.6, METEOR = 24.6
    - MP-LSTM (VGGNet): Bleu@4 = 34.8, METEOR = 24.8

- MP-LSTM (C3D+VGGNet): Bleu@4 = 35.8, METEOR = 25.3
- SA (GoogleNet): Bleu@4 = 35.2, METEOR = 25.2
- SA (VGGNet): Bleu@4 = 35.6, METEOR = 25.4
- SA (C3D): Bleu@4 = 36.1, METEOR = 25.7
- SA (C3D+VGGNet): Bleu@4 = 36.6, METEOR = 25.9
- aLSTMs: Bleu@4 = 38.0, METEOR = 26.1

VI. Acknowledgements:

The mentioned work received support from the following funding sources in China:

1. National Natural Science Foundation of China:
    a. Project Number: 61502080
    b. Project Number: 61632007
2. Fundamental Research Funds for the Central Universities:
    a. Project Number: ZYGX2016J085
    b. Project Number: ZYGX2014Z007

References:

Among the total 52 references, here are some commonalities and patterns:

- *Topic Focus*: The references primarily revolve around topics related to computer vision, machine learning, deep learning, natural language processing, and their applications in tasks such as image and video understanding, captioning, and annotation.
- *Machine Learning Approaches*: Many references discuss various machine learning approaches, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), long short-term memory (LSTM), and their variants. These approaches are applied for tasks like image and video captioning, translation, recognition, and retrieval.
- *Datasets and Benchmarks*: Several references mention specific datasets or benchmarks used for evaluation, such as MSR-VTT, as a standard for assessing image and video captioning methods.
- *Semantic Understanding*: There is a focus on understanding and extracting semantics from images and videos, including techniques like attention

mechanisms, semantic embeddings, and multimodal learning to bridge the gap between visual content and natural language descriptions.

- *Evaluation Metrics*: Common evaluation metrics such as BLEU, METEOR, and other automatic evaluation methods are referenced, highlighting the importance of assessing the quality of generated captions.
- *Incorporation of Text and Vision*: Many references emphasize integrating textual and visual information, highlighting the interplay between natural language and visual content, especially in tasks like image and video captioning.
- *Applications*: References discuss applications like image and video captioning, video concept learning, image retrieval, and related areas, indicating a strong interest in real-world applications of the discussed methodologies.

---x---x---x---