KIPS Trans. Softw. and Data Eng.
Vol.13, No.1 pp.35~49

https://doi.org/10.3745/KTSDE.2024.13.1.35
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

# Analysis of Research Trends in Deep Learning-Based Video Captioning

Lyu Zhi† · Eunju Lee†† · Youngsoo Kim†††

## ABSTRACT

Video captioning technology, as a significant outcome of the integration between computer vision and natural language processing, has emerged as a key research direction in the field of artificial intelligence. This technology aims to achieve automatic understanding and language expression of video content, enabling computers to transform visual information in videos into textual form. This paper provides an initial analysis of the research trends in deep learning-based video captioning and categorizes them into four main groups: CNN-RNN-based Model, RNN-RNN-based Model, Multimodal-based Model, and Transformer-based Model, and explain the concept of each video captioning model. The features, pros and cons were discussed. This paper lists commonly used datasets and performance evaluation methods in the video captioning field. The dataset encompasses diverse domains and scenarios, offering extensive resources for the training and validation of video captioning models. The model performance evaluation method mentions major evaluation indicators and provides practical references for researchers to evaluate model performance from various angles. Finally, as future research tasks for video captioning, there are major challenges that need to be continuously improved, such as maintaining temporal consistency and accurate description of dynamic scenes, which increase the complexity in real-world applications, and new tasks that need to be studied are presented such as temporal relationship modeling and multimodal data integration.

Keywords : Video Captioning, Computer Vision, Natural Language Processing, Deep Learning

# 딥러닝 기반 비디오 캡셔닝의 연구동향 분석

려 치† · 이 은 주†† · 김 영 수†††

## 요 약

컴퓨터 비전과 자연어 처리의 융합의 중요한 결과로서 비디오 캡셔닝은 인공지능 분야의 핵심 연구 방향이다. 이 기술은 비디오 콘텐츠의 자동 이해와 언어 표현을 가능하게 함으로써, 컴퓨터가 비디오의 시각적 정보를 텍스트 형태로 변환한다. 본 논문에서는 딥러닝 기반 비디오 캡셔닝의 연구 동향을 초기 분석하여 CNN-RNN 기반 모델, RNN-RNN 기반 모델, Multimodal 기반 모델, 그리고 Transformer 기반 모델이라는 네 가지 주요 범주로 나누어 각각의 비디오 캡셔닝 모델의 개념과 특징 그리고 장단점을 논하였다. 그리고 이 논문은 비디오 캡셔닝 분야에서 일반적으로 자주 사용되는 데이터 집합과 성능 평가방안을 나열하였다. 데이터 세트는 다양한 도메인과 시나리오를 포괄하여 비디오 캡션 모델의 훈련 및 검증을 위한 광범위한 리소스를 제공한다. 모델 성능 평가방안에서는 주요한 평가 지표를 언급하며, 모델의 성능을 다양한 각도에서 평가할 수 있도록 연구자들에게 실질적인 참조를 제공한다. 마지막으로 비디오 캡셔닝에 대한 향후 연구과제로서 실제 응용 프로그램에서의 복잡성을 증가시키는 시간 일관성 유지 및 동적 장면의 정확한 서술과 같이 지속해서 개선해야 할 주요 도전과제와 시간 관계 모델링 및 다중 모달 데이터 통합과 같이 새롭게 연구되어야 하는 과제를 제시하였다.

키워드 : 비디오 캡션, 컴퓨터 비전, 자연어 처리, 딥러닝

## 1. Introduction

Video captioning is a complex task aimed at automatically generating descriptive and informative textual descriptions. This process involves a deep understanding of the visual content of videos and converting it into natural language expressions. Due to the specificity of video data, it is necessary to understand the objects, attributes, actions [1-3], and events in the video, as well as establish temporal and semantic contextual relationships [4-8]. Common video captioning applications encountered in daily life include: providing access to videos for visually impaired and hearing-impaired individuals, real-time descriptions of surveillance footage, and retrieval of online video content [9].

With the emergence of excellent deep learning models, significant progress has been made in the development of temporal relationships, semantic coherence, and contex-

※ 이 논문은 2022년 전주대학교 연구비 지원을 받아 수행되었음.
† 준 회 원 : 전주대학교 문화기술학과 박사수료
†† 비 회 원 : 전주대학교 문화기술학과 박사과정
††† 정 회 원 : 전주대학교 인공지능학과 교수
Manuscript Received : October 5, 2023
First Revision : November 17, 2023
Accepted : November 28, 2023
* Corresponding Author : Youngsoo Kim(pineland@jj.ac.kr)

tual understanding in videos, which has driven the advancement of video captioning technology. However, the field of video captioning still faces challenges.

This paper provides a systematic analysis of the applications of deep learning in video caption generation and points out future research directions. To be more specific, we first explore the basic principles and pros and cons of several representative methods in video caption generation in recent years. Next, we summarize commonly used datasets and representative evaluation metrics in video captioning. Finally, we discuss other issues in existing research, challenges in video captioning, and future research directions.

Chapter 2 provides a comprehensive overview of the latest advancements in video captioning technologies. It delves into various research studies, methodologies, and approaches employed in the field. In Chapter 3, datasets commonly used for video captioning evaluation, along with the associated evaluation metrics, are introduced. The challenges faced in video captioning are thoroughly discussed in Chapter 4, followed by the proposal of future research directions to address these challenges. Finally, Chapter 5 comprehensively summarizes the key findings and contributions of this article, highlighting the main points and insights derived from the research.

## 2. Approaches to Video Captioning Research

This section aims to review various methods utilized in the field of video captioning, primarily focusing on four distinct strategies based on deep learning: CNN-RNN, RNN-RNN, Transformer-based, and Multimodal approaches. Fig. 1 shows the four methods of video captioning in the captioning generation challenge. By providing detailed explanations of these approaches, it highlights the continuous development of video captioning technology. CNN (Convolution Neural Network) have revolutionized image tasks, including object detection and image classification. RNN(Recurrent Neural Network) excel in processing sequential data for natural language tasks, aiding in machine translation and text classification. As a result, these advancements have provided powerful tools for language modeling, machine translation, and more. The innovation of video captioning technology benefits from the development of deep learning. By conducting in-depth research on different methods, a better understanding of the development in this field can be achieved, offering support for future research and practice.

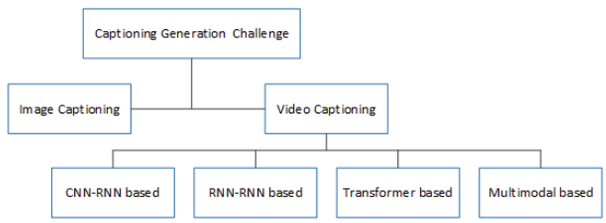



Fig. 1. An overview of Captioning Generation Problem

### 2.1 CNN-RNN based Model

The CNN-RNN model is a well-established architecture in video captioning, primarily achieving the task of translating video content into text by combining the visual processing capability of CNN with the sequence generation capability of RNN. As shown in Fig. 2, The CNN and RNN respectively play the roles of encoder and decoder. The output of the encoder serves as input to the decoder, which generates captions based on the encoded information.

In 2014, Venugopalan et al. [10] introduced a deep learning model for end-to-end translation of video2text. They employed a pre-trained CNN to extract image features, transformed them into fixed-length representations, and then fed these representations into a two-layer LSTM to decode them into word sequences, generating text. The advantage of this method is the elimination of the need for traditional multi-step processing, simplifying the process. However, this approach has limitations in adequately considering temporal information in video processing, which may constrain its performance on longer videos.

In 2015, Li Yao et al. [4] proposed a method that combines 3D CNN (3D Convolutional Networks) to capture temporal characteristics of dynamic representations and utilizes a pre-trained GoogleNet[11] to extract spatial features. They also selected Histograms of HoG, HoF and MbH(Oriented Gradients, Oriented Flow, and Motion Boundary) to accurately extract motion features of local tem-

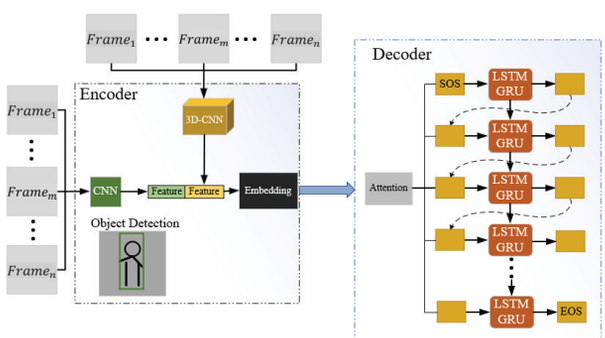



Fig. 2. Video Captioning Tasks based on CNN-RNN Model

poral structures, reducing the computational cost of the 3D CNN. They introduced temporal attention mechanism based on soft attention to selectively attend to and process key temporal vectors, which were then input into an LSTM-based decoder to produce captions.

In 2017, Gan Z et al. [5] introduced the SCN(Semantic Compositional Network). They emphasized that incorporating semantics can lead to improved video captioning generation with interpretability. The model employed ResNet [12] and C3D(Convolutional 3D) [13] to extract features from images and videos, detected semantics to determine the probabilities of respective labels, and input semantic features and visual features into the LSTM(Long Short Term Memory) [14] to generate captions.

In 2020, Chen H et al. [6] pointed out several issues in previous video captioning works, including the lack of meaningful semantic features, performance differences among different generation strategies, and inadequate representation of video content. They proposed metrics to measure meaningful features and used them as input to the SDN(Semantic Detection Network). In the encoding stage, they extracted video frame and video features using ResNeXt [15] and ECO(Efficient convolutional network) [16], respectively, concatenated the two features, and obtained Semantic Features. During the decoding stage, SDN-enhanced LSTM was implemented to effectively capture the temporal information of the video. This approach mitigates the issues of gradient vanishing or exploding, resulting in descriptions that are more closely aligned with the video content.

In 2021, Perez-Martin J et al. [17] introduced the SemSynAN(Visual-Semantic-Syntactic Aligned Network), which leverages combinations of visual, semantic, and syntactic representations to generate superior captions from the decoder. In the encoding stage, they used CNN and 3D-CNN were employed to extract features from both individual video frames and videos. They then employed concept detectors to retrieve relevant keywords and obtain semantic representations. Finally, Visual-Syntactic Embedding is used to map visual features into a common space with word labels for alignment processing. In the decoding stage, they employed three dedicated RNN layers: v-se-LSTM (Visual-Semantic Layer), v-sy-LSTM (Visual-Syntactic Layer), and se-sy-LSTM (Semantic-Syntactic Layer), along with time attention mechanisms and integration gates, to accurately retrieve information from the visual and semantic syntax-related layers.

In 2022, Yan et al. [18] proposed the GL-RG (Global-local representation granularity) to capture global-local representations within video frames sufficiently. The model consists of three parts in the encoder to model different video scopes. The Long-range Encoder employs CNN to build context features and uses 3D CNN to capture global temporal correspondences, guiding vocabulary generation. The Short-range Encoder captures motion using CNN and 3D-Resnet18 [19]. The Local-keyframe Encoder learns local semantic vocabulary using a residual network and generates integrated features through linear layers. In the decoding stage, LSTM [14] is used to convert integrated features into word sequences for caption generation.

Table 1 shows the models, datasets and evaluation results used in the CNN-RNN-based model. Although the

Table 1. Models, Datasets, and Evaluation Results used in CNN-RNN based Video Captioning Methods

| | MODEL | Dataset | BLUE-4/% | METEOR/% | CIDEr/% | ROUGE-L/% |
|---|---|---|---|---|---|---|
| 2014, Venugopalan et al. [10] | CNN+LSTM [14] | MSVD [47] | 30.77 | 27.66 | - | - |
| 2015, Li Yao et al. [4] | 3D CNN+LSTM [14] | Youtube2Text (MSVD) [47] | 41.92 | 29.6 | 51.67 | - |
| Gan Z et al. [5] | ResNet [12] and C3D [13]+LSTM [14] | Youtube2Text (MSVD) [47] | 51.10 | 33.50 | 77.70 | - |
| | | Charades [50] | 14.50 | 18.40 | 23.70 | - |
| 2020, Chen H et al. [6] | ResNeXt [15] and ECO [16] + S-LSTM | Youtube2Text (MSVD) [47] | 62.40 | 39.00 | 109.70 | 77.00 |
| | | MSR-VTT [48] | 45.80 | 29.30 | 53.20 | 63.60 |
| 2021, Perez-Martin J et al. [17] | CNN and 3D CNN+ v-se-LSTM v-sy-LSTM and se-sy-LSTM | MSVD [47] | 64.40 | 41.90 | 111.50 | 79.50 |
| | | MSR-VTT [48] | 46.40 | 30.40 | 51.90 | 64.70 |
| 2022, Yan et al. [18] | CNN and 3D CNN+LSTM [14] | MSR-VTT [48] | 46.90 | 30.40 | 55.00 | 63.90 |
| | | MSVD [47] | 57.70 | 38.60 | 95.90 | 74.90 |

CNN-RNN model is simpler and performs better compared to traditional models, there are still some challenges:

Firstly, using only CNN or 3D CNN for visual processing fails to capture the temporal relationships of the entire video, thereby leading to a deficiency in video context. Secondly, as the sequence length increases, RNN may fail to capture distant information.

### 2.2 RNN-RNN based Model

The RNN-RNN model bears similarities to the seq2seq model. As shown in Fig. 3, two different RNNs are used as an encoder and a decoder. The primary distinction between the RNN-RNN model and the CNN-RNN model lies in the fact that the former commonly utilizes a CNN+RNN architecture as its encoder. This approach enables better modeling of video data, which is based on temporal sequences.

In 2015, Venugopalan et al. [9] introduced the S2VT (Sequence to sequence-video to text), which innovatively utilized the Sequence-to-Sequence model for video description purposes. For feature extraction, they used VGG-16 [20] to extract features from video frames, which were then fed as inputs to an LSTM [14] to learn sequence representations of the frames. Additionally, they improved object activity classification using Optical Flow [21]. Two-layer LSTMs were used in both the encoder and decoder, with the first layer handling video feature input and the second layer constructing hidden representations of both text and video sequences for text generation.

In 2015, Donahue et al. [22] proposed the LRCN (Long-term recurrent convolutional networks), which combines CNNs for image feature extraction and LSTMs for video caption generation. CNNs were used to extract image features, and these features were then provided as input to an LSTM, which in turn served as the input to the decoder LSTM responsible for generating video descriptions.

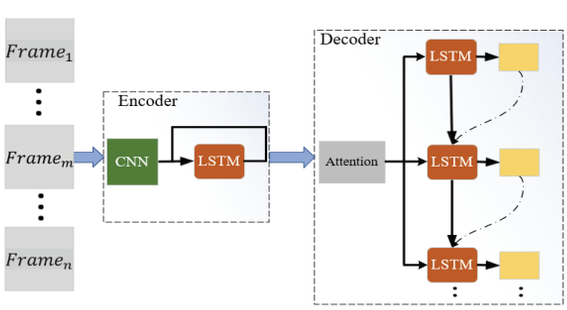



Fig. 3. Video Captioning Tasks based on RNN-RNN Model

In 2016, Pan et al. [23] introduced the HRNE (Hierarchical recurrent neural encoder), an encoder-decoder architecture. They utilized GoogleNet [11] for extracting frame-level features and used linear embedding of these features as model inputs. In the encoding phase, a two-layer LSTM processed the input information, with the first LSTM performing modeling at a local time scale and the second LSTM capturing long-term dependencies in the video sequence. To capture local temporal structure more effectively, they designed a receptive field similar to a convolutional neural network, which improved the encoding of video sequences and provided robust semantic phase to receive encoder information and generate captions.

In 2017, Gao et al. [24] introduced aLSTM (attention-based LSTM). They used Inception-v3 [25] to extract video frame features during the encoding phase and utilized an LSTM to obtain a holistic understanding of the video. In the decoding phase, they incorporated an attention mechanism to leverage contextual information and selectively retrieve crucial semantic information, resulting in more accurate and meaningful descriptions compared to previous methods.

In 2018, Wang, Ma, Zhang, et al. [7] proposed the RecNet (Reconstruction Network), which represented a departure from previous approaches. The model introduces the concept of a 'Reconstructor' to generate text consistent with video content, going beyond mere content-based generation. It reconstructs the overall structure of the video through Mean Pooling applied to the hidden state sequence produced by the decoder, addressing both local and global structures. The Reconstructing Local Structure component used soft attention on the decoder's hidden state sequence to select key hidden states, improving the reconstruction of local structural information in video frames. RecNet incorporated a bidirectional learning mechanism for both forward (video2text) and reverse (text2video) flows, thereby enhancing video captioning tasks.

In 2020, Zhang and Peng [1] introduced the OSTG (object-aware spatio-temporal graph), considering object recognition and spatiotemporal correlation and aggregation. OSTG consisted of three main parts: global context encoder, the temporal relation encoder, and spatial relation encoder. The temporal relation encoder learned spatiotemporal paths of object regions. Using VLAD (Vector of Locally Aggregated Descriptors) [26], local features of

object regions were extracted, and ConvGRUs (Convolutional Gated Recurrent Unit) captured spatiotemporal features of object regions. The spatial relation encoder encoded spatial interactions between different objects. It built spatial relation graphs based on visual feature similarity, spatial overlap, and center point distances among objects, applying GCN(Graph Convolutional Networks) [27] to model inter-object relationships. The global context encoder captured the global context by aggregating features of global frames with VLAD and integrating local features into global VLAD representations. During the decoding phase, considering the information from three encoders, hierarchical attention and temporal attention are used to determine the importance of each temporal step in the region. The encoded information from the encoders is fed into two GRUs (Gated Recurrent Units) [59] to generate captions.

Table 2 shows the models, datasets and evaluation results used in the RNN-RNN-based model. Although the RNN-RNN model can effectively consider the temporal information of videos and model the dynamic changes and event sequences, the increase in sequence length can lead to gradient vanishing or explosion, which affects the quality of generated descriptions. Furthermore, it lacks complex modeling capabilities for scenes with deep semantic content.

### 2.3 Transformer based Model

Fig. 4 presents an approach based on the Transformer model, which utilizes its encoder and decoder components. Typically, CNN models are used as input for the encoder. The decoder then generates more accurate, vivid and coherent subtitles.

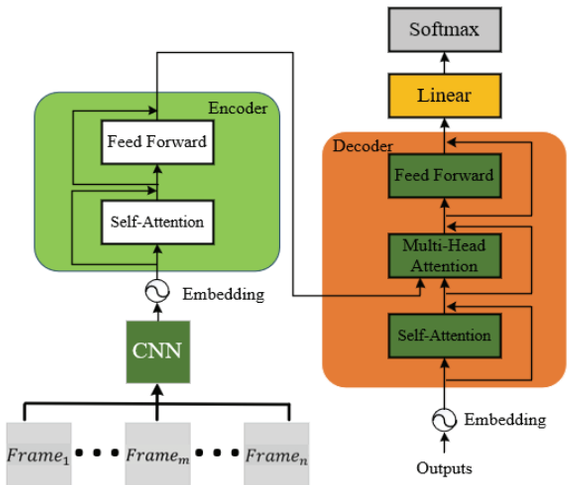



Fig. 4. Video Captioning Tasks based on Transformer

In 2018, Zhou L et al. [28] proposed a model to address the problem of separately training temporal proposals and caption models. This model consists of Video Encoder Stack, Proposal Decoder, and Caption Decoder. It utilizes attention to encode consecutive frames for visual features. The Proposal Decoder employs a TCN (Temporal Convolutional Network) to generate event proposals from the visual features of the video encoder. Finally, the event proposals and visual features are fed into the Caption Decoder, facilitating the generation of video captions.

In 2020, Pan et al. [2] introduced a graph-based model that effectively captures object interactions in both spatial and temporal dimensions. They use an object-aware knowledge refinement mechanism to enhance caption generation performance. This model represents objects as nodes and their relationships as edges. In the encoding phase, they extract scene and object features from frames

Table 2. Models, Datasets, and Evaluation Results used in RNN-RNN based Video Captioning Methods

| | MODEL | Dataset | BLUE-4/% | METEOR/% | CIDEr/% | ROUGE-L/% |
|---|---|---|---|---|---|---|
| 2015, Venugopalan S et al. [9] | VGG-16 [20]+ and LSTM [14]+LSTM [14] | MSVD [47] | - | 29.20 | - | - |
| | | MPII-MD [54] | - | 7.10 | - | - |
| | | M-VAD [55] | - | 6.70 | - | - |
| 2015, Donahue J et al. [22] | CNN and LSTM+Two layer LSTM | TACoS multi-level [61] | 28.80 | - | - | - |
| 2016, Pan P et al. [23] | HRNE+LSTM [14] | MSVD [47] | 43.80 | 6.80 | - | - |
| 2017, Gao L et al. [24] | Inception-v3 [25] and LSTM [14] + aLSTM | MSVD [47] | 50.80 | 33.30 | 74.80 | - |
| | | MSR-VTT [48] | 38.00 | 26.10 | 43.20 | - |
| 2018, Wang B et al. [7] | Inception-V4 [62] and LSTM [14] + SA-LSTM | MSVD [47] | 52.30 | 34.10 | 80.30 | 69.80 |
| 2020, Zhang J, Peng Y. [21] | object detection, C-GRU and GCN [27]+GRU [59] | MSVD [47] | 57.50 | 36.80 | 92.10 | - |
| | | MSR-VTT [48] | 41.90 | 28.60 | 48.20 | - |

using ResNet-101 [12] and Faster R-CNN [29]. Object features are transformed into an adjacency matrix for graph representation, and a GCN [27] updates node features. In the decoding phase, they use the Transformer to train scenes and objects individually and introduce object-aware knowledge refinement to integrate different feature spaces, thereby generating video captions.

In 2021, Wang T et al. introduced the PDVC (video captioning with parallel decoding) [30] framework, a Transformer-based approach that achieves end-to-end dense video captioning using parallel decoding. They use a pre-trained model in the encoding phase and maintain sequence information between frames through projection using linear layers and positional encoding. During the parallel decoding phase, an event counter is utilized to accurately determine the occurrence of events in the video, facilitating the generation of complete and consistent video captions. The Localization head uses event query features to accurately predict event boundaries, providing precise timing information for caption generation. The Captioning head allows the direct generation of detailed captions for each event, enabling end-to-end dense video captioning.

In 2022, to address issues like redundant connections, smooth transitions, and relationship ambiguity within the Transformer framework, Li L et al. [3] proposed the LSTG (long short-term graph). The relationships between objects in terms of their spatial and temporal dimensions can be effectively captured by utilizing the STG (Short-Term Graph) and LTG (Long-Term Graph) methodologies. STG builds short-term spatial semantic relationships between objects for adjacent frames, considering relative positions and appearance similarities. In order to capture long-term dependencies and transformations between objects, LTG constructs connections among objects using a sparse approach. The LSRT(Long Short-Term Relation Transformer) module replaces the traditional self-attention mechanism [39] in the Transformer with a G3RM (Global Gated Graph Reasoning Module), which improves caption generation by providing accurate object positioning and temporal information.

In 2022, Ye H et al. [8] introduced the HMN (Hierarchical Modular Network), consisting of Entity Module, Predicate Module, and Sentence Module, which take a detailed approach to caption generation. The Entity Module maps input objects to representations using a Transformer encoder-decoder architecture, generating semantic embeddings related to video content. The Predicate Module takes as input object features related to video motion and uses an attention mechanism and Bi-LSTM to encode motion. The Sentence Module integrates initial video context features, motion features, and object features to generate captions by utilizing attention mechanisms for aggregating relevant information and summarizing global video context. The Description Generator uses LSTM to generate captions step by step, considering three levels: video representations, language predictions, and previous words.

Table 3 shows the datasets and evaluation results used in the Transformer-based model. Despite the satisfactory performance of Transformer-based methods in video captioning generation, they encounter several challenges. One of the challenges is processing the temporal sequence information in videos, as videos consist of a collection of frames that involve temporal evolution. Furthermore, it is necessary to conduct further research on effective integration methods of image features with text and audio information. Additionally, generating natural language descriptions that align with the content of the video is of utmost importance.

Table 3. Datasets, and Evaluation Results used in Transformer based Video Captioning Methods

| | Dataset | BLUE-4/% | METEOR/% | CIDEr/% | ROUGE-L/% |
|---|---|---|---|---|---|
| 2018, Zhou L et al. [28] | ActivityNet Captions [49] | 2.77 | 11.11 | - | - |
| 2020, Pan B et al. [2] | MSVD [47] | 52.20 | 36.90 | 93.00 | 73.90 |
| | MSR-VTT [48] | 40.50 | 28.30 | 47.10 | 60.90 |
| 2021, Wang T et al. [30] | ActivityNet Captions [49] | 1.96 | 8.08 | 28.59 | - |
| | YouCook2 [53] | 0.80 | 4.70 | 22.71 | - |
| 2022, Li L et al. [3] | MSVD [47] | 55.60 | 37.10 | 98.50 | 73.50 |
| | MSR-VTT [49] | 42.60 | 28.30 | 49.50 | 61.00 |
| 2022, Ye H et al. [8] | MSVD [47] | 59.20 | 37.70 | 104.00 | 75.10 |
| | MSR-VTT [49] | 43.50 | 29.00 | 51.50 | 62.70 |

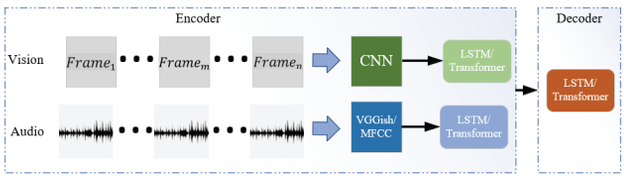



Fig. 5. Video Captioning Tasks based on Multimodal Model

### 2.4 Multimodal based Model

Fig. 5 is a model based on a multimodal approach that mainly combines multiple feature inputs, including video and audio in the video, to better understand the video content, and fuses these features through the encoder to generate accurate and coherent subtitle descriptions.

In 2016, Ramanishka et al. [31] proposed MMVD (Multimodal Video Description), an extension of S2VT [9]. This model extracts features from 26 frames sampled at intervals, including ResNet [12], C3D [13], and MFCC (Mel Frequency Cepstral Coefficients) [32]. Simultaneously, it represents video category information using one-hot encoding. The visual and audio features extracted are embedded into a low-dimensional space to facilitate feature fusion. These fused features are then encoded using LSTM [14] and subsequently decoded to generate video descriptions.

In 2019, Xu et al. [33] introduced the Semantic-Filtered Soft-Split-Aware Gated LSTM, which integrates semantic information with audio features to enhance the quality of captioning. Visual features of the video are obtained through ResNet [12] and 3D-ResNext, while audio features are augmented and appended to keyframe features through audio augmentation. The encoding phase consists of the SSAG-LSTM-E (Soft Split Recognition Gate Multilayer LSTM Encoder), the SF-LSTM-E (Semantic Filtering LSTM Video Encoder) for capturing video features of the sequence, and the SSAG-LSTM-E for capturing video features of the sequence. These features and SSAG-LSTM-E are used for processing videos with multiple segments.

In 2019, Sun et al. [34] proposed VideoBERT (Video Bidirectional Encoder Representations from Transformers), aiming to address the data scarcity issue in video captioning and overcome the limitations of fixed-length captions using a hierarchical vector quantization-based method. VideoBERT extends the BERT (Bidirectional Encoder Representations from Transformers) [35] and can learn advanced features from both video and language. In the video processing phase, it extracts visual features using S3D (separable 3D CNN) [36], labels these features using hierarchical k-means, and retrieves text from YouTube's ASR (automatic speech recognition) automatically. VideoBERT initializes with pre-training from the BERTLARGE model, takes visual and text features as input, and completes the final task.

In 2020, Ging S et al. proposed COOT (Cooperative Hierarchical Transformer) [37], which models different levels and modes. This model comprises Attention-aware Feature Aggregation and Contextual Transformer. Attention-aware feature aggregation methods extract local visual and semantic features from video and text, using attention mechanisms to interlink these features. The Contextual Transformer, with multiple transformer layers, learns both high-level and low-level information, ensuring semantic consistency across modes. It's employed to integrate local and global contexts in video content.

In 2020, Iashin and Rahtu [38] proposed the MDVC (Multi-modal Dense Video Captioning module), which utilizes ASR to obtain temporally aligned textual descriptions of audio and uses them as separate inputs alongside video frames. This module adopts the Transformer [39] as a template and comprises two core modules: the Captioning Model and the Temporal Event Localization Model. The Temporal Event Localization Model is used to generate time intervals where events may occur in the video and uses the Bi-SST (Bidirectional Single-stream Temporal Action Proposal Network) for this purpose. The Captioning Model generates descriptions for event proposals, utilizing I3D (Inflated 3D ConvNet) [40] for visual features and VGGish [41] for audio features. These features, along with previous time-step word embeddings, are input into the Transformer [39], and its output combines all modalities and estimates the probability distribution of the vocabulary.

In 2022, Luo et al. [42] proposed Clip4clip to address the problem of insufficient visual representation in video captioning. Clip4clip mainly consists of three modules. The first module is Video Encoder, that employs the VIT-B/32 model [43] to encode videos, dividing them into image blocks and forming interactions between these image blocks through a Transformer [39] to obtain the final representation. The second module is Text Encoder, that directly utilizes the text encoder of the CLIP (Contrastive Language-Image Pre-training) [44]. The third module is Similarity Calculation part, that is divided into three types: Sequential type, parameter-free type and Tight type. The parameter-free type does not consider the temporal order

Table 4. Models, Datasets, and Evaluation Results used in Multimodal based Video Captioning Methods

| | MODEL | Dataset | BLUE-4/% | METEOR/% | CIDEr/% | ROUGE-L/% |
|---|---|---|---|---|---|---|
| 2016, Ramanishka V et al. [31] | ResNet [12], C3D [13], MFCC [32] and LSTM [14] +LSTM [14] | MSR-VTT [48] | 40.70 | 28.60 | 46.50 | 61.00 |
| 2019, Xu Y et al. [33] | ResNet [12], C3D [13], Audio Feature and SSAG-LSTM +LSTM [14] | MSR-VTT [48] | 40.80 | 28.70 | 46.80 | 61.50 |
| 2019, Sun C, Myers A, Vondrick C, et al. [34] | S3D [36], ASR and BERT [35] and Transformer [39] | YouCook2 [53] | 4.33 | 11.94 | 0.55 | 28.80 |
| 2020, Ging S et al. [37] | BERT[35], Resnet-152[12], 3D ResNext-101+ Transformer [39] | YouCook2 [53] | 11.30 | 19.85 | 57.24 | 37.94 |
| | | ActivityNet Captions [49] | 10.85 | 15.99 | 28.19 | 31.45 |
| 2020, Iashin V, Rahtu E. [38] | C3D [13], VGGish [41] and I3D [40]ASR + Transformer [39] | ActivityNet Captions [49] | 2.86 | 11.72 | - | - |
| 2022, Luo H et al. [42] | VIT-B/32 [43], CLIP [44], LSTM [14] and + Transformer [39] | MSR-VTT [48] | 49.80 | 31.40 | 59.70 | 65.70 |
| | | MSVD [46] | 55.90 | 36.90 | 122.60 | 73.90 |
| 2022, Seo P H et al. [45] | BERT [35] and ViViT [46] + Transformer [39] | MSR-VTT [47] | 48.92 | 38.66 | 0.60 | 64.00 |
| | | ActivityNet Captions [49] | 6.84 | 12.31 | - | - |
| | | YouCook2 [53] | 21.88 | 27.09 | 2.21 | 49.38 |

of frames and directly computes the similarity between video and text using average pooling of features from all frames. The Sequential type models video frame sequences with LSTM [14] or Transformer [39] encoders, captures temporal order between frames, and generates a semantic representation of the entire video, followed by computing the similarity between this representation and the text. The Tight model utilizes a Transformer [39] encoder for facilitating multimodal interaction between video and text. It predicts similarity through the use of linear layers, predicting similarity through linear layers.

In 2022, Seo and Ji [45] proposed a new generative pretraining framework called MV-GPT (Multimodal Video Generative Pretraining). In the encoding phase, the Text Encoder uses BERT [35] to extract semantic features from text input and learn contextual relationships between words. On the other hand, the Visual Encoder utilizes the ViViT (Video Vision Transformer) from MV-GPT [46] as the video encoder. This allows the direct extraction of visual features from raw pixels, enabling end-to-end training and flexibility. These two encoders fuse information through a co-attentional transformer module, allowing for context and attention interactions between text and visual features. In the decoding phase, the decoder uses a self-attention mechanism [39] to process inputs and better capture dependencies and contextual information between different modalities. Additionally, the decoder enhances the model's generation capability through Masked Language Modeling (MLM) loss based on the current generated word.

Table 4 shows the models, datasets and evaluation results used in the Multimodal-based model. Although multimodal caption generation can better understand the fusion of visual and textual information, aligning the modalities during fusion needs to be considered, as well as how to effectively integrate the modalities. Sometimes, the same event may be presented differently across different modalities, and when fusing modalities, it is necessary to consider the balance of weights between different modalities. Even in videos, there are cases where the shapes of objects change before and after multiple events, making it difficult to accurately connect them and infer the captions effectively.

## 3. Datasets and Evaluation for Video Captioning

### 3.1 Popular Datasets for Video Captioning

In this section, we will explain and introduce in detail some commonly used datasets. These datasets cover various scenarios, including daily activities, movies, cooking, etc. The main method of data collection is to gather YouTube videos or movie clips.

1) MSVD: Microsoft Research Video Description Corpus

MSVD [47] is a dataset for video captioning research,

which was introduced by the University of Texas at Austin and Microsoft Research in 2011. This dataset comprises over 1,970 distinct video clips, covering various topics, scenes, and activities.

2) MSR-VTT: Microsoft Research Video to Text

MSR-VTT [48] is a comprehensive video dataset designed specifically for video-to-text tasks. It consists of 10,000 video clips with a total duration of 41.2 hours. The dataset includes 200,000 sentence-sentence pairs, making it one of the most extensive collections in terms of both sentences and words.

3) ActivityNet Captions

ActivityNet Captions [49] is used for video understanding and subtitle generation tasks. The dataset consists of 20,000 untrimmed YouTube videos with an average length of 120 seconds. Each video is accompanied by three or more manually produced sentence captions, averaging 13.5 words per video.

4) Charades

Charades dataset [50] offers 9,849 daily indoor activity videos with an average length of 30 seconds. It includes interactions between 46 object categories in 25 indoor scenes, as well as 157 action categories composed of 30 verbs.

5) VATEX

VATEX [51] is a large-scale multilingual and multimodal dataset that includes 41,250 videos and 825,000 English-Chinese caption pairs. The dataset provides videos and corresponding textual descriptions, encompassing over 206,000 translation pairs.

6) LSMDC: Large Scale Movie Description Challenge

LSMDC [52] provides a large-scale dataset consisting of 118,081 segments extracted from 202 movies. The segment captions are derived from scripts or descriptive video services for people with visual impairments.

7) YouCook2

YouCook2 [53] is a large cooking video dataset from YouTube, featuring 2,000 unedited third-person perspective videos. Covering 89 global recipes, it offers about 22 videos per recipe.

8) MPII-MD

MPII-MD [54] dataset contains about 68,000 video segments from 94 Hollywood movies, each with a brief single-sentence description sourced from scripts, audio descriptions, or descriptive video service (DVS).

9) M-VAD: Montreal Video Annotation Dataset

M-VAD [55] is a dataset composed of over 24,000 video segments obtained via a semi-automatic method. It includes 63,000 annotations (face bounding boxes) related to character appearances, as well as 34,000 text annotations associated with them.

### 3.2 Evaluation Metrics for Video Captioning

In this section, we will provide a detailed explanation and introduction of commonly used evaluation metrics. These metrics serve as widely accepted tools for assessing the quality of generated video captions. Their purpose is to quantitatively measure the similarity and accuracy between automatically generated captions and human reference captions, enabling an objective assessment of model performance.

1) BLEU: Bilingual Evaluation Understudy

BLEU [56] is a measurement standard used to automatically evaluate machine translation quality. It measures the degree of agreement between candidate translations (machine translation results) and reference translations (human translation results). The key idea is based on n-gram (contiguous n words) accuracy. The final BLEU score is calculated by obtaining weighted averages of the number of matching n-grams between candidate translations and reference translations. When calculating n-gram accuracy, BLEU also considers the modified n-gram accuracy to address the issue of weighted averaging of n-grams of different lengths. To account for translation length, BLEU introduced a penalty factor for short sentences. The BLEU score applies a penalty to longer candidate translations compared to reference translations, addressing length discrepancies. It uses weighted averages for the significance of different-length n-grams in evaluation. The weights or importance factors are calculated as the geometric mean of the logarithms of n-gram accuracies, reflecting the relative importance of n-grams of different lengths in the evaluation. Finally, BLEU calculates the score as a geometric mean of the weighted average of n-gram accuracy and the penalty factor for short

sentences, resulting in a score between 0 and 1. A higher score indicates better machine translation quality. The formula is as follows:

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

Here, BP (Brevity Penalty) is a penalty factor that adjusts the length difference between candidate translations and reference translations. $w_n$ is a weight coefficient used to calculate the weighted average of n-gram accuracy. $p_n$ represents n-gram accuracy, which indicates the ratio of correctly matching n-grams in candidate translations to the total number of n-grams in candidate translations.

When calculating accuracy, considering modified n-gram accuracy, it takes the logarithm of n-gram accuracy values and calculates the weighted average. This reflects the relative importance of n-grams of different lengths in the evaluation. The penalty factor BP adjusts the length of candidate translations to ensure that candidate translations match reference translations in terms of length. The calculation of BP is as follows:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1-r/c) & \text{if } c \le r \end{cases} \quad (2)$$

Among them, r represents the length of the candidate translation, and c represents the length of the reference translation that is closest to r. By utilizing the aforementioned formula, we can obtain a BLEU score ranging from 0 to 1, with a higher score indicating superior translation quality by the machine.

2) METEOR: Metric for Evaluation of Translation with Explicit ORdering

METEOR [57] is an improvement upon BLEU, aiming to address inherent deficiencies in the BLEU standards. METEOR employs word matching based on surface forms, stem forms, and semantics. The METEOR scoring involves Fmean and Penalty.

Fmean is used to comprehensively evaluate the word precision and recall in the METEOR metric, measuring the accuracy and coverage of translations by calculating their harmonic mean. The formula for calculating Fmean is as follows:

$$F_{mean} = \frac{\text{precisio} * \text{recall}}{(1-\alpha) * recall + \alpha * \text{precisio}} \quad (3)$$

METEOR considers the case of longer matches by segmenting the matching words in the system translation into chunks, then calculating a penalty term based on the number of chunks and the number of matched chunks. The formula is as follows:

$$Penalty = \gamma * \left(\frac{chunks}{unigrams_{matched}}\right)^{\beta} \quad (4)$$

Here, $\gamma$ is a penalty factor used to balance the strength of the penalty term. $\frac{chunks}{unigrams_{matched}}$ represents the number of words matched in the system translation that are divided into chunks. $\beta$ is an exponential parameter that determines the relationship between the number of matching chunks and word matches. Finally, the calculation formula for METEOR is as follows:

$$METEOR = F_{mean} * (1 - Penalty) \quad (5)$$

3) CIDEr: Consensus-based Image Description Evaluation

CIDEr [58] is an evaluation metric commonly utilized for image captioning tasks to assess the similarity between generated and reference sentences. This metric primarily operates by calculating the Term Frequency-Inverse Document Frequency (TF-IDF) vectors for the n-grams present in each sentence. The formula used for TF-IDF calculation is as follows:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right) \quad (6)$$

Here, $\Omega$ represents the set of all n-grams, $h_k(S_{ij})$ denotes the count of phrase $w_l$ appearing in the reference sentence $s_{ij}$, $|I|$ is the total number of images in the dataset, and cosine similarity is employed to measure the semantic consistency between candidate and reference sentences. The following formula is used for this calculation:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g_n(c_i) \cdot g_n(s_{ij})}{\|g_n(c_i)\| \|g_n(s_{ij})\|} \quad (7)$$

Here, $c_i$ represents the candidate sentence, $S_i$ is the set of reference sentences, m is the number of reference sen-

tences, n is the length of n-gram, $g_n(c_i)$ and $g_n(s_{ij})$ represent the TF-IDF vectors of the candidate and reference sentences.

4) ROUGE: Recall-Oriented Understudy for Gisting Evaluation)

ROUGE [59] is a metric for automatically evaluating summary quality. ROUGE [59] is a metric for automatically evaluating summary quality. Its function is to quantify the quality of summaries by comparing the overlap between machine-generated summaries and reference summaries. ROUGE includes multiple different measurement methods, such as ROUGE-N, ROUGE-L, etc., which can be used to evaluate the matching of summaries at different levels. The "N" in ROUGE-N refers to N-gram. The calculation method of ROUGE-N is similar to BLEU, but it focuses on calculating the recall of N-grams. For N-gram, the ROUGE-N score can be calculated. The formula is as follows:

$$ROUGE-N=\frac{\Sigma_{S\in\{ReferemceSummaries\}}\sum_{gram\in S}Count_{match}(gram_n)}{\Sigma_{S\in\{ReferemceSummaries\}}\sum_{gram\in S}Count(gram_n)} \tag{8}$$

Here, $n$ represents the length of n-gram, $gram_n$represents n-gram unit. The numerator quantifies the N-grams shared between the reference translation and the machine translation, while the denominator calculates the total count of N-grams in the reference translation.

ROUGE-L primarily focuses on Longest Common Subsequence (LCS), where L stands for Longest. ROUGE-L measures the structural and sequential similarity between generated translation C and reference translation S. The calculation formula is as follows:

$$R_{LCS}=\frac{LCS(C,S)}{\mathrm{len(S)}} \tag{9}$$

$$P_{LCS}=\frac{LCS(C,S)}{\mathrm{len(C)}} \tag{10}$$

$$F_{LCS}=\frac{\left(1+\beta^2\right)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2P_{LCS}} \tag{11}$$

Here, $R_{LCS}$ represents the recall rate, $P_{LCS}$ represents the precision rate, $\beta$ represents the balance parameter that trades off recall and precision.

## 4. Future Works and Challenges

In the field of video captioning, significant progress has been made, but there are still various challenges that must be addressed and overcome. These challenges are crucial for generating captions that maintain temporal consistency and accurately capture dynamic scenes, which are essential for ensuring the quality and accuracy of captions.

Firstly, ensuring temporal consistency presents one of the primary challenges in generating video captions. Videos encompass various scenes, actions, and dialogue changes in a dynamic multimedia format. In this scenario, generating captions that are consistent with the video context and maintain temporal consistency is a complex task. To address the challenge of temporal consistency, multiple factors need to be considered. One approach is to ensure that the generated captions align with the video context. To maintain both language expression and temporal consistency with dynamic changes in the video, analyzing the video to recognize key scenes and action transition points and inserting or updating captions at appropriate times may be necessary. Another approach is to use attention mechanisms to handle temporal consistency. Attention mechanisms dynamically weigh frames at different time steps during the decoding process to maintain the temporal sequence of the generated captions with respect to the video content. Attention mechanism dynamically weights the frames at different time steps during the decoding process to maintain the temporal order of the generated captions relative to the video content. By considering the relevance of video frames, the model can make more accurate decisions on when to generate or update captions, thereby enhancing overall caption-to-video consistency.

Secondly, accurately capturing dynamic scenes in caption generation is also a complex problem involving various complexities. When processing videos where multiple individuals, objects, and backgrounds interact, a caption generation system needs to have a high level of observation to accurately and meticulously capture each dynamic element and express them clearly and vividly in language. In such situations, merely understanding the literal meaning is insufficient. Deeply interpreting visual information is required to grasp the emotions, actions, and intentions inherent in interactions. Dynamic scenes may require specific identification and descriptions in captions, allowing the audience to understand complex situations accurately.

In order to address these challenges, researchers continuously explore innovative technical paths. The rapid development of deep learning has provided powerful tools for enhancing caption generation techniques. Specifically, we categorize them as:

1) **Precisely model the visual content of the video and its temporal relationships.** Enhance the recognition and understanding of objects, actions, and scenes within the video, and create more sophisticated models for the relationships between objects in the time series.

2) **Integrate visual information with data from other modalities (e.g. audio, text).** Perform information fusion across modalities, align and analyze multimodal data, and leverage the complementarity of multimodal data to enhance the quality and effectiveness of video analysis.

3) **Perform context modeling.** Video is a unique multimedia form, composed of multiple continuous images arranged in chronological order. Therefore, it is imperative to delve into enhancing the modeling of contextual information within the video to ensure improved coherence and consistency of video descriptions.

4) **Perform long-range dependency modeling.** In some videos, events unfold over an extended period, requiring the modeling of long-range dependencies to generate accurate descriptions. Traditional RNN structures may be limited in modeling long-range dependencies, so researchers can explore more effective architectures.

5) **Consider datasets and evaluation metrics.** To comprehensively assess the quality and temporal consistency of generated captions, it is crucial to construct richer multimodal datasets and design more accurate evaluation metrics.

In future research, it is important to consider the interplay between objects and actions during the encoding stage. Video frames that feature multiple objects exhibit behaviors and actions that are intricately linked to the corresponding video captions. Moreover, the inseparable relationship between audio and events in videos enables the possibility of event analysis through audio for video processing. Therefore, exploring the incorporation of multimodal techniques in video caption generation represents a crucial avenue for future research.

## 5. Conclusion

This paper provides a comprehensive overview of four types of video captioning models: CNN-RNN, RNN-RNN, multimodal, and Transformer models. It emphasizes their unique advantages and applications. Additionally, this paper introduces well-known video caption datasets and explains key evaluation metrics for video captions to facilitate an objective understanding of video caption model performance evaluation.

Furthermore, the paper engages in an in-depth discussion of several challenges faced by the video captioning field. Maintaining temporal consistency is a critical challenge, considering that videos consist of a continuous series of frames, and ensuring the seamless generation of captions requires the consideration of continuous context information transfer and processing. Additionally, accurately capturing dynamic scenes is another challenging task. Videos contain various movements, object interactions, and scene changes, and models need to precisely recognize and describe these dynamic changes to generate accurate captions.

The paper also proposes future research directions and challenges in video captioning. Firstly, improving the ability to model video content and temporal relationships to better understand the meaning and emotional changes in videos. Secondly, the integration of visual information with other mode data (e.g., sound and text) offers opportunities to enhance the accuracy and diversity of caption generation. Moreover, the paper provides more detailed mentions of research directions and challenges in video captioning to offer researchers more specific guidelines.

## References

[1] J. Zhang and Y. Peng, "Video captioning with object-aware spatio-temporal correlation and aggregation," *IEEE Transactions on Image Processing*, Vol.29, pp.6209-6222, 2020.

[2] B. Pan et al., "Spatio-temporal graph for video captioning with knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10870-10879, 2020.

[3] L. Li, X. Gao, J. Deng, Y. Tu, Z. Zha, and Q. Huang, "Long short-term relation transformer with global gating for video captioning," *IEEE Transactions on Image Processing*, Vol.31, pp.2726-2738, 2022.

[4] L. Yao et al., "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE International Conference on Computer Vision*, pp.4507-4515, 2015.

[5] Z. Gan et al., "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5630-5639, 2017.

[6] H. Chen, K. Lin, A. Maye, J. Li, and X. Hu, "A semantics-assisted video captioning model trained with scheduled sampling," *Frontiers in Robotics and AI*, Vol.7, pp.475767, 2020.

[7] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7622-7631, 2018.

[8] H. Ye, G. Li, Y. Qi, S. Wang, Q. Huang, and M. Yang, "Hierarchical modular network for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.17939-17948, 2022.

[9] 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*, pp.4534-4542, 2015.

[10] S. Venugopalan et al., "Translating videos to natural language using deep recurrent neural networks[J]," *arXiv preprint arXiv:1412.4729*, 2014.

[11] C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015.

[12] 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*, pp.770-778, 2016.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international Conference on Computer Vision*, pp.4489-4497, 2015.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol.9, No.8, pp.1735-1780, 1997.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1492-1500, 2017.

[16] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proceedings of the European Conference on Computer Vision ECCV*, pp.695-712, 2018.

[17] J. P. Martin, B. Bustos, and J. Pérez, "Improving video captioning with temporal composition of a visual-syntactic embedding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.3039-3049, 2021.

[18] L. Yan et al., "Gl-rg: Global-local representation granularity for video captioning," *arXiv preprint arXiv:2205.10706*, 2022.

[19] D. Tran, J. Ray, Z. Shou, S. F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, Vol.17, No.1-3, pp.185-203, 1981.

[22] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2625-2634, 2015.

[23] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1029-1038, 2016.

[24] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, Vol.19, No.9, pp.2045-2055, 2017.

[25] 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*, pp.2818-2826, 2016.

[26] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp.3304-3311, 2010.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[28] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.8739-8748, 2018.

[29] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp.1440-

1448, 2015.

[30] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.6847-6857, 2021.

[31] V. Ramanishka et al., "Multimodal video description," in *Proceedings of the 24th ACM International Conference on Multimedia*, pp.1092-1096, 2016.

[32] B. Logan et al., "Mel frequency cepstral coefficients for music modeling," in *Ismir. Plymouth*, MA Vol.270, p.11, 2000.

[33] Y. Xu, J. Yang, and K. Mao, "Semantic-filtered soft-split-aware video captioning with audio-augmented feature," *Neurocomputing*, Vol.357, pp.24-35, 2019.

[34] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.7464-7473, 2019.

[35] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[36] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision* ECCV, pp.305-321, 2018.

[37] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "Coot: Cooperative hierarchical transformer for video-text representation learning," Advances in *Neural Information Processing Systems*, Vol.33, pp.22605-22618, 2020.

[38] V. Iashin and E. Rahtu, "Multi-modal dense video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.958-959, 2020.

[39] A. Vaswani et al., "Attention is all you need," Advances in *Neural Information Processing Systems*, Vol.30, 2017.

[40] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.6299-6308, 2017.

[41] S. Hershey et al., "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.131-135, 2017.

[42] H. Luo et al., "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, Vol.508, pp.293-304, 2022.

[43] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[44] A. Radford et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, pp.8748-8763, 2021.

[45] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.17959-17968, 2022.

[46] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić and C. Schmid, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ICCV), pp.6836-6846, 2021.

[47] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp.190-200, 2011.

[48] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5288-5296, 2016.

[49] R. Krishna, K. Hata, F. Ren, F-F. Li, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp.706-715, 2017.

[50] G. A. Sigurdsson, G. Varol, X. L. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam*, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer, pp.510-526, 2016.

[51] X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.4581-4591, 2019.

[52] A. Rohrbach et al., "Movie description," *arXiv preprint*, 2016.

[53] L. Zhou, C. Xu, and J. J Corso, "Towards automatic learning of procedures from web instructional videos," in

*AAAI Conference on Artificial Intelligence*, pp.7590-7598, 2018.

[54] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3202-3212, 2015.

[55] S. Pini, M. Cornia, F. Bolelli, L. Baraldi, and R. Cucchiara, "M-vad names: a dataset for video captioning with naming," *Multimedia Tools and Applications*, Vol.78, pp.14007-14027, 2019.

[56] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.311-318, 2002.

[57] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp.65-72, 2005.

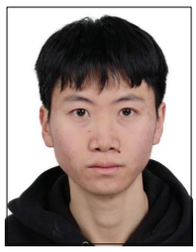
[58] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4566-4575, 2015.

[59] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, pp.74-81, 2004.

[60] K. Cho. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[61] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, "Coherent multi-sentence video description with variable level of detail," in *Pattern Recognition: 36th German Conference*, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36. Springer, pp.184-195, 2014.

[62] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.31. No.1, 2017.



**LYU ZHI**

https://orcid.org/0009-0000-3048-4032

e-mail : lyuzhi@jj.ac.kr

He received his Master's degree in agricultural informatics from Shanxi Agricultural University, China, in 2017. Since 2019, he has studied artificial intelligence in the Department of Culture and Technology at Jeonju University. His research interests include deep learning and video captioning.

**Eunju Lee**

https://orcid.org/0009-0006-3741-3775

e-mail : leeeunju@jj.ac.kr

She received a Master's degree in Smart Agro ICT Convergence from Jeonju Univ. in 2022. She is currently pursuing a Ph.D. degree in the Artificial Intelligence major at the Department of Cultural Technology at Jeonju Univ. Her research interests include deep learning and image processing.

**Youngsoo Kim**

https://orcid.org/0000-0002-6214-7222

e-mail : pineland@jj.ac.kr

He received a B.S. degree in computer science in 1994 from the Republic of Korea Air Force Academy, an M.S. degree in computer science engineering from Sogang University in 2001, and a Ph.D. degree in computer science engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 2009. He worked as a researcher in the Battlefield Informatization Lab of the Military Development Research Center of the Korea National Defense Research Institute (KIDA) until 2021. He has been an assistant professor in the Department of Artificial Intelligence, Jeonju University, Korea. His research interests include artificial intelligence, IoT, cloud computing, and edge/fog computing.