Improving Dense Video Captioning with a Transformer-based Multimodal Fusion Model

Yixuan Liu, Ziwei Zhou*, Shuyue Hui, Haoyuan Ma, Hongju Li, Zhibo Zhang
School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China
**Corresponding Author*

Abstract: Dense Video Captioning (DVC) plays a pivotal role in advancing video understanding within computer vision and natural language processing. Traditional DVC models have predominantly focused on visual information, often neglecting the auditory component. To address this limitation, we propose a Transformer-based multimodal fusion model that integrates audio and visual cues for comprehensive multimodal input processing. Built on an encoder-decoder architecture, the model synergizes audio and visual streams. The feature encoder combines self-attention mechanisms with convolutional neural networks to achieve precise audio feature encoding, while the decoder employs multimodal fusion by leveraging intermodal confidence scores to adaptively integrate inputs. A feedforward neural network enhances historical textual representations, and strategic skip connections eliminate redundant data, prioritizing key video features for refined captioning. Extensive validation on benchmark datasets MSR-VTT and MSVD demonstrates that our model outperforms existing methods, achieving BLEU-4, ROUGE, METEOR, and CIDEr scores of 0.427, 0.618, 0.294, and 0.532 on MSR-VTT, and 0.539, 0.741, 0.369, and 0.976 on MSVD. By effectively leveraging the complementary strengths of audio and visual data, our model establishes a new benchmark in DVC, offering precise and comprehensive video content interpretation.

Keywords: Audio-Visual Integration; Dense Video Captioning; Multimodal Fusion; Transformer Networks

1. Introduction

The rapid advancement of internet technologies has resulted in an exponential increase in video data, presenting significant challenges in the efficient parsing and comprehension of vast content volumes. Consequently, automating video content analysis and description has emerged as a crucial area of research and technological development. This trend has heightened interest in Dense Video Captioning technology, emphasizing the necessity for enhanced methods to interpret and manage the continuously expanding volume of video content.

Zhou et al. [1] introduced a groundbreaking approach in this field by employing a Transformer-based framework to develop an end-to-end model for dense video description. This model was notable for its innovative use of a new masking scheme, which enabled efficient simultaneous training of the event detection and description phases. Building on this work, Wang et al. [2] further advanced the field by proposing a fully end-to-end framework with parallel decoding capabilities. This enhancement aimed to improve task collaboration and reduce the reliance on manual parameters during event detection. Additionally, Wang et al. [3] combined Video Transformer (ViT) with deep semantic learning to create a model that optimizes feature extraction networks like ResNet152 and C3D through deep separable convolution techniques. This integration was designed to lower computational overhead and reduce model complexity in video description tasks.

Iashin et al. [4] emphasized the importance of multimodal features in videos through their innovative approach. Their methodology focused on the comprehensive integration of features derived from various pre-trained models to capture video content across visual, audio, and textual dimensions. Specifically, they used the I3D convolutional network for

visual feature extraction, the VGGish network for audio analysis, and an automatic speech recognition system to convert speech content into textual descriptions. This approach led to the development of a more enriched and precise framework for enhancing video comprehension.

Li et al. [5] introduced the innovative MSTVC model, which employs the R(2+1)D network for precise visual feature extraction and utilizes a semantic detector to generate key semantic information. Additionally, they integrated audio features and introduced a multiscale deformable attention mechanism, complemented by a parallel prediction head strategy. This combination resulted in a substantial acceleration of model convergence.

Mun et al. [6] efficiently enhanced the generation of coherent text descriptions by implementing a two-tier reward system and conducting a comprehensive analysis of video context. In a related study, Chen et al. [7] introduced a weak supervision technique focused on multi-instance concept learning to strengthen the integration of event detection and description tasks. This approach improves the model's understanding of inherent video features by incorporating an induced aggregate attention mechanism. Additionally, it leverages identified keywords during feature alignment to produce more precise text descriptions, thereby fostering a stronger relationship between the two tasks.

This paper presents the Transformer-based Video Feature Model, a novel multimodal fusion framework for dense video description that enhances the ability to capture both local and global dependencies in audio and video features. Our proposed methodology represents a significant advancement in improving video content understanding and description generation by incorporating a self-attention mechanism-based feature encoder, a multimodal fusion decoder, and a feedforward neural network (FNN) to optimize historical text features.

2. Related Theorie

Transformer networks[8] have revolutionized natural language processing (NLP) with their distinctive architecture, which is centered on self-attention mechanisms and offers a more efficient approach to processing sequential data. Unlike earlier models that relied on recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, Transformers excel in managing long-distance dependencies by capturing global relationships within input sequences solely through self-attention mechanisms.

The Transformer network comprises multi-head attention mechanisms, positional encoding, feedforward neural networks, and layer normalization components at its core. These elements work synergistically to achieve exceptional performance across diverse domains beyond NLP, including computer vision, speech recognition, and related fields. Moreover, the Transformer ensures stable information propagation within deep network structures, a feature that is critical for generating precise video descriptions. By integrating these advanced components, the Transformer substantially enhances the efficiency of video description generation models, underscoring its immense potential in natural language processing and beyond.

3. Model Design

Dense video description, a subfield of video parsing, is distinguished by its focus on extended video durations and diverse content involving multiple events. The input for such tasks consists of a sequence of time-ordered video frames $v = \{v_t\}$, $t\epsilon 0, \cdots, T-1$, while the output comprises a series of sub y_i from the set sub y, where each sub $y = \left(y^{start}, y^{end}, \{v_j\}\right)$, and a sequence of vocabulary terms v_j that form individual sentences, drawn from the vocabulary library V and varying in length. TInitially, the model performs event segmentation to generate a set of events $P = \left\{t_i^{start}, t_i^{end}, score_i, h_i\right\}$, with the $score$ denoting the event's rating, and h_i serving as input for subsequent video description tasks, ultimately producing the final descriptive text. Due to the inherent complexity of dense video descriptions, previous research has often approached these tasks in two distinct phases. However, this methodology frequently led to a disconnect between the generated descriptions and an over-reliance on the precision of event segmentation.

In this study, a novel Transformer-based model, referred to as the Transformer-based Video Feature Model, is proposed, utilizing an

encoder-decoder architecture. As illustrated in Fig. the model employs an attention mechanism to capture dynamic interactions among frames, between different events, and between events and frames within the video. This approach enables the construction of a comprehensive set of feature representations for event queries.

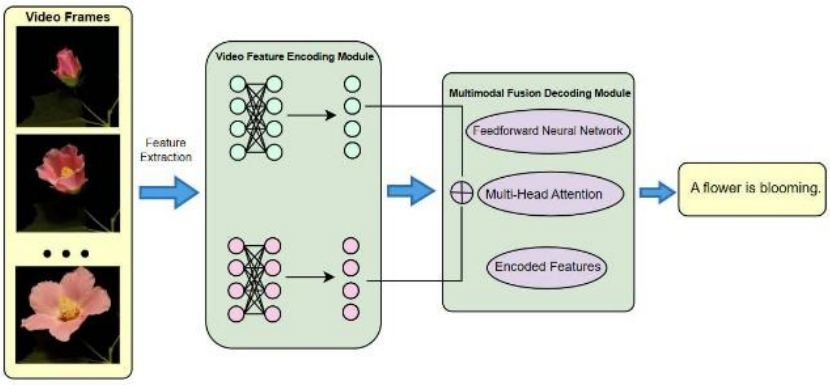

Figure 1. Overall Framework of the Proposed Transformer-based Video Feature Model

Furthermore, the model incorporates two parallel predictors that simultaneously estimate the boundaries and descriptions of each event query. The number of events, denoted as N_{set}, is determined using an event counter mechanism from a global perspective. During the output phase, the top N_{set} events with the highest confidence levels are selected, ensuring the coherence and completeness of the generated narrative.

3.1 Feature Extraction

This article presents a model that integrates feature extraction from three distinct modalities: video, audio, and text.

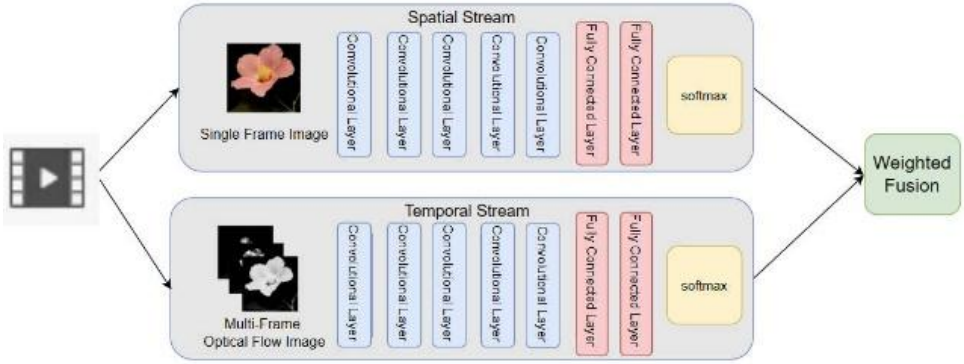

Figure 2. Two-Stream Network Architecture Diagram

(1) Recent research predominantly employs 3D convolutional neural networks (3D CNNs) or two-stream networks to extract visual information from videos, as these methods have demonstrated remarkable performance in capturing video features. The 3D CNN thoroughly analyzes video frames, leveraging its deep learning architecture to capture rich spatio-temporal information[10]. Its 3D convolutional kernel processes not only the visual content of individual frames but also the temporal relationships between consecutive frames.

The two-stream network, illustrated in Fig. 2, is specifically designed to enhance the understanding of video behavior. The process begins by sampling RGB frames at 25 frames per second and capturing optical flow frames using the PWC-Net model[16]. The frames are resized to a minimum side length of 256 pixels, after which a 244×244 block is cropped from the center. These pre-processed frames are subsequently input into the pre-trained C3D model, which extracts spatio-temporal features over 2.56-second segments (equivalent to 64 frames) of the original video, generating a high-dimensional feature vector. This sequence of operations ensures that the extracted visual features encapsulate comprehensive spatio-temporal dynamics and motion information.

(2) Audio Features: Capturing audio features is critical for constructing the model. Audio signals can be represented in various domains, such as time, frequency, and cepstral[11]. with

the time and frequency domains being commonly preferred for their practicality. Once converted into the frequency domain, the audio signal is input into a neural network system for advanced feature extraction.

The extraction process begins by separating the audio signal from the video and resampling it to a single-channel signal at 16 kHz. To optimize the model's performance and reduce audio information loss, pre-emphasis processing is applied to enhance the signal's high-frequency components and balance the spectrum. The audio signal is then segmented into approximately stationary small segments using a Hamming window of 25 ms, with framing and window function processing applied at 15 ms steps. Next, the segments undergo Fourier transformation to convert the signal into the frequency domain, followed by conversion into the power spectrum using a 64-stage MEL filter at MEL frequency. The final output is a 96×64 logarithmic MEL spectrogram, which is then input into a pre-trained VGG network for further feature extraction. Each spectrogram produces a d_a -dimensional feature vector, where $A \epsilon R^{n_a \times d_a}$, with n_a representing the number of audio segments in the video.

Neural networks significantly enhance the system's ability to learn abstract concepts from audio signals, improving processing efficiency and accuracy in audio-related tasks.

(3) Neural networks cannot inherently understand natural language text[13]; thus, text content must first be converted into word vectors for computational recognition.

3.2 Encoding Module

This model processes inputs from two modalities using specialized encoders. When managing multi-modal inputs, it is crucial to account for the inherent differences in data representation across modalities, emphasizing the necessity of targeted encoding structures. Utilizing a single encoder for multi-modal inputs may result in incomplete feature extraction or the loss of critical information.

The visual feature encoder consists of N layers[14], where the first layer receives visual features as input and generates intermediate representations. Each subsequent layer uses the output of the previous layer as its input. Each encoding layer comprises two sub-layers: a multi-head attention layer and a feedforward layer. These sub-layers are connected via residual connections and normalized using layer normalization, thereby forming the structure of the visual feature encoder.

The visual feature V_s is input into the visual encoder for visual feature interaction. The input visual features are denoted as $V \epsilon R^{n_v \times n_d}$, where n_v visual features and d_v represents their dimensionality. The visual feature encoder updates the visual features at the sss-th encoding layer according to the following process:

$$\widetilde{V}_s = MultiHead[LayerN(V_s)] \quad (1)$$

$$V_{s+1} = \widetilde{V}_s + FNN\left[LayerNorm(\widetilde{V}_s)\right] \quad (2)$$

$$FNN(V) = FC\{Drop\left[ReLU\left(FC(V)\right)\right]\} \quad (3)$$

The visual feature encoder integrates multiple layers, each with distinct components. The feedforward network layer, denoted as $FFN(\,)$ performs transformations to extract high-level features. The dropout layer, $Drop(\,)$ mitigates overfitting by randomly deactivating a subset of neurons during training. The activation function, $ReLU(\,)$, introduces non-linear relationships between neural layers, enabling the model to capture complex patterns. The multi-head attention layer, $MultiHead(\,)$, forms the core of the encoder by leveraging self-attention mechanisms to enhance stability and robustness in feature representation.

(2) The audio feature encoder combines a convolutional neural network (CNN) with a Transformer to capture both local and global audio features. The Transformer effectively models global audio context, while the CNN focuses on local audio details. This hybrid structure includes a feedforward network layer, a multi-head attention layer, a convolutional layer, and a second feedforward network layer.

The multi-head attention layer primarily enhances the identification of correlations across various audio segments, facilitating the capture of overall audio characteristics. However, it is less effective at modeling intricate local features. To address this, the convolutional layer, positioned after the multi-head attention layer, strengthens the integration of local and global audio features by capturing local correlations and effectively learning localized information.

The convolutional layer reduces computational costs and the number of parameters by employing pointwise convolution and one-dimensional depthwise

convolution, thereby mitigating the risk of overfitting[12]. This approach effectively models relationships between audio signals and enhances the network's expressive power when integrated with other activation functions.

The input audio features are defined as $A \epsilon R^{n_a \times d_a}$, where n_a represents the total number of audio segments after frame segmentation, and d_a denotes the dimensionality of each audio feature. At the s-th encoding layer, the audio feature A_s is input into the multi-head attention mechanism to model global dependencies within the audio data. The specific process is as follows:

$$\widetilde{A_s} = \overline{A_s} + MultiHead(\overline{A_s}, \overline{A_s}, \overline{A_s}) \quad (4)$$

$$\overline{A_s} = A_s + \frac{1}{2} FFN(A_s) \quad (5)$$

Subsequently, the modeled audio features are processed through the convolutional layer to capture local information, resulting in the final output of encoded audio features.

$$\widetilde{\widetilde{A_s}} = \widetilde{A_s} + Conv(\widetilde{A_s}) \quad (6)$$

$$A_{s+1} = Layernorm\left[\widetilde{\widetilde{A_s}} + \frac{1}{2} FFN\left(\widetilde{\widetilde{A_s}}\right)\right] \quad (7)$$

where $Conv(\,)$ represents the convolutional layer.

Using the confidence scoring system, the relative weight and significance of each modal feature in the text generation task are effectively measured and compared. This module introduces a confidence threshold mechanism, which functions as an advanced filter to accurately identify and select critical modal features essential for the current text generation task. By targeting and retaining these key features, information from other modal features is utilized to iteratively optimize the less significant ones. This process achieves a complementary integration of information across modalities, enhancing the model's capacity to flexibly utilize multimodal contextual information while effectively addressing the problem of audiovisual redundancy.

After the feature fusion step, the integrated key features are passed into the gating layer. The algorithm calculates the scale factor for each modal feature, analyzing and quantifying its detailed impact on the final text generation. This ensures that reasonable weights are assigned to different modal features in the text generation task. The implementation details are outlined in Equations (8) through (10).

$$G_s = \delta \cdot V_s^{\alpha} + (1 - \delta) A_s^{\alpha} \quad (8)$$

$$\delta = Sigmoid\left[FC(VA)\right] \quad (9)$$

$$VA_s = G_s + FNN(G_s) \quad (10)$$

Finally, the integrated features are passed into a fully connected layer, where a softmax function is applied to compute the probability distribution for the next word in the generated text. The specific calculation is detailed in Equation (11).

$$P = f_{\text{Softmax}}\left[FNN(VA)\right] \quad (11)$$

To address potential interference caused by historical text features, the decoding side introduces two parallel processing paths: a redundant discard path and an information enhancement path. This structure highlights features relevant to the current context while minimizing the influence of historical information.

4. Experiment and Analysis

Author names and affiliations are to be centered beneath the title and printed in Times New Roman 11-point, non-boldface type. (See example below)

4.1 Dataset

In the video description task, although there are multiple thematic datasets, such as those related to film or food, this study focuses on two widely used general datasets: Microsoft Video Description (MSVD)[13] and Microsoft Research-Video to Text (MSR-VTT)[20]. An overview of the specific information on these two datasets is provided below.

The MSVD dataset includes an average of 41 individual statements per video, while the MSR-VTT dataset has 20 individual statements per video.

Table 1. Dataset Details

Data set	Total Videos	Average Duration	Common divisions	Total Vocabulary	Unique Words
MSVD	1970	10sec	1200/100/670	607399	13010
MSR-VTT	10000	20sec	6513/497/2900	1856523	29316

4.2 Sample Data Preprocessing

The audio data was divided into 0.96-second segments and organized in a time series. A Hamming window function was applied with a window length of 25 ms and a sliding interval of 15 ms. Each segment

was processed with the window function, followed by a Fourier transform to convert it into the frequency domain. Subsequently, a MEL filter generated the spectrogram, and the pre-trained VGG network embedded the audio features into a 128-dimensional vector space.

We sampled frames from the video data at a rate of 25 frames per second. The visual features of the video were extracted using the trained C3D network, with a down-sampling ratio of 4, resulting in each time step consisting of four frames. This produced a 1024-dimensional visual feature representation, where each feature encapsulates 2.56 seconds of information from the original video. For the text data, we used the pre-trained GloVe model to map each word into a vocabulary space with 10,172 dimensions, resulting in a 10,172-dimensional word vector for each word. This provides the foundation for subsequent processing and analysis of the text information.

4.3 Model Training

The experimental environment is based on a 64-bit Ubuntu 22.04 LTS operating system. The central processing unit is an Intel Pentium G4060 running at 3.50 GHz, with 16 GB of physical memory and 16 GB of virtual memory. The graphics card used is an RTX 3060 Ultra OC model. All experiments were conducted using the PyTorch framework. During the feature extraction stage, we set a batch size of 16 and directly processed the uncropped visual and audio features. To ensure consistent input feature dimensions, the visual and audio features were extended to lengths of 300 and 800, respectively, adequately covering all feature lengths in the training set..

The batch size in the text generation module is 32, and it processes visual and audio features extracted from event video segments based on the reference time period. To maintain a consistent sequence length for different modal features within the same batch, each modal feature is padded to match the longest sequence length in the batch. Various modal features, each possibly having distinct dimensions, are standardized into a 1024-dimensional internal space for computational purposes. Each multi-head attention layer consists of 4 layers (N) and 4 subspace heads (H). The confidence threshold (α) for the audio-visual attention layer in the multimodal fusion decoding module is set at 0.15.

To enhance rapid convergence and prevent overfitting during training, the model parameters are optimized using the Adam optimizer. Weight decay is employed to manage model complexity, with a decay rate of 0.0001. The learning rate is set to 0.00001, the total number of iterations is 70, and the dropout rate is 0.2. Early stopping is triggered to terminate the training if there is no improvement in text generation on the validation set for 30 consecutive epochs..

In our experiments on the MSR-VTT dataset, we compared the performance of the Baseline model and the audio-visual fusion model using the CIDEr evaluation metric, as depicted in Fig. 3. The analysis shows that the fusion model exhibits superior performance by effectively combining visual and audio features. Visual features convey information about scenes, objects, and actions, while audio features enhance the environmental and contextual aspects of sound. By integrating these two modalities, the model's semantic parsing ability is significantly improved, leading to a richer and more accurate understanding.

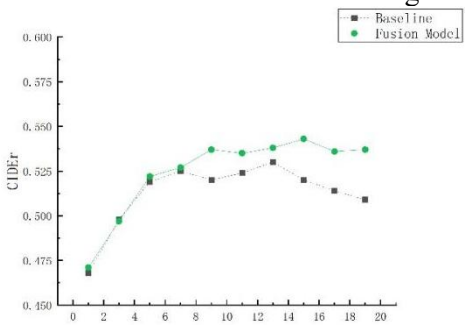

Figure 3. Comparison of Changes in CIDEr Indicators

4.4 Comparative Experiment and Analysis

To assess the model's performance, various encoder-decoder structural models were selected for comparison using the MSVD and MSR-VTT datasets. The detailed results are presented in Tables 2 and 3.

Table 2. Comparative Results of Different Methods on MSVD Dataset

MODEL	B_4	ROUGE	METEOR	CIDER
TTA	0.517	0.718	0.337	0.876
CSA-SR	0.525	0.723	0.356	0.835
POS-CG	0.517	0.703	0.341	0.918
SAAT	0.467	0.693	0.321	0.81
RMN	0.542	0.732	0.367	0.942
SGN	0.528	0.736	0.355	0.943

STG-KD	0.52	0.734	0.369	0.920
Our method	**0.539**	**0.741**	**0.369**	**0.976**

Table 3. Comparative Results of Different Methods on MSR-VTT Dataset

MODEL	B_4	ROUGE	METEOR	CIDER
TTA	0.413	0.617	0.276	0.465
CSA-SR	0.414	0.619	0.283	0.497
POS-CG	0.413	0.617	0.284	0.476
SAAT	0.405	0.607	0.281	0.477
RMN	0.417	0.61	0.283	0.487
SGN	0.413	0.608	0.287	0.487
STG-KD	0.401	0.609	0.287	0.467
Our method	**0.427**	**0.618**	**0.294**	**0.532**

The data in Table 2 demonstrates that this model performs better than other models on multiple performance indicators in the MSVD dataset. In the CIDEr index, the model showed a notable 3.5% enhancement. The CIDEr index, designed for assessing descriptive texts, closely aligns with human evaluation criteria, demonstrating the superiority of this model. This confirms that incorporating attention mechanisms and guidance signals can greatly improve the model's performance.

The model also showed improvements in performance across all evaluation indicators on the MSR-VTT dataset, as presented in Table 3. The BLEU_4 score rose by 3.39%, while the ROUGE score increased by 1.64%, the METEOR score by 0.7%, and the CIDEr score significantly by 9.24%. The significant rise in the CIDEr score highlights the strong performance of fusion networks and contextual semantic capture models with attention mechanisms in generating descriptions.

4.4 Experimental Results

At the conclusion of the experiment, two videos were randomly selected from the test set to assess the model's effectiveness, as depicted in Fig 4 and Fig 5.

Baseline:Some people are skating.
The Model in This Paper:Five people are having a skating match.

Figure 4. Experimental Results (a)

Baseline: Some ducks are swimming.
The Model in This Paper: Eleven ducks are swimming together.

Figure 5. Experimental Results (b)

By analyzing the experimental data above, it is evident that the model accurately aligns the subject, target object, and action in the sentence. These experimental results further confirm that integrating attention mechanisms and guidance signals significantly enhances the model's ability to maintain semantic consistency. The model effectively indicates the number of subjects participating in the activity, such as "five people" and "eleven ducks."

5. Conclusion

This chapter introduces a dense video description method that utilizes multimodal fusion. A Transformer encoder, enhanced by convolutional techniques, is employed to effectively model the local and global dependencies of audio features. An audio-visual attention module is proposed for feature fusion. Additionally, the feature representation of historical text is enhanced through the use of a feedforward neural network (FNN). Skip connections are integrated to construct a redundant discard path and an information enhancement path, leveraging subtraction and addition operations for this purpose. Building upon the Transformer model, the inclusion of a mechanism to select historical information enhances the diversity of feedforward processes, making it more suitable for applications in video description, video search, and retrieval. By effectively utilizing historical text features, the model places greater emphasis on the current input information, ultimately improving its generalization ability. Ablation experiments were conducted to assess the impact of these methods on the model's performance. The model was then compared with existing mainstream models using classical datasets, conclusively demonstrating its effectiveness. The model's impact on video description is also visually demonstrated.

Acknowledgments

This paper is supported by University of Science and Technology Liaoning College student innovation and entrepreneurship project: Blind assistance system based on visual recognition and NLP (NO.S202410146037).

References

[1] L. Zhou, Y. Zhou, J. J. Corso, et al., "End-to-end dense video captioning with masked transformer," IAENG International Journal of Computer Science,

vol. 2018, no. 023, pp. 8739-8748, 2018.
[2] J. Wang, W. Jiang, L. Ma, et al., "Bidirectional attentive fusion with context gating for dense video captioning," IAENG International Journal of Computer Science, vol. 2018, no. 023, pp. 7190-7198, 2018.
[3] Y. Xiong, B. Dai, D. Lin, "Move forward and tell: a progressive generator of video descriptions," IAENG International Journal of Computer Science, vol. 2018, no. 023, pp. 468-483, 2018.
[4] J. Mun, L. Yang, Z. Ren, et al., "Streamlined dense video captioning," IAENG International Journal of Computer Science, vol. 2019, no. 023, pp. 6588-6597, 2019.
[5] Y. Li, T. Yao, Y. Pan, et al., "Jointly localizing and describing events for dense video captioning," IAENG International Journal of Computer Science, vol. 2018, no. 023, pp. 7492-7500, 2018.
[6] C. Y. Ma, A. Kadav, I. Melvin, et al., "Attend and interact: higher-order object interactions for video understanding," IAENG International Journal of Computer Science, vol. 2018, no. 023, pp. 6790-6800, 2018.
[7] H. Xu, B. Li, V. Ramanishka, et al., "Joint event detection and description in continuous video streams," IAENG International Journal of Computer Science, vol. 2019, no. 023, pp. 396-405, 2019.
[8] J. Wang, S. Zeng, W. Li, et al., "Attention Mechanism Video Description Model Based on Dilated Convolution," Electronic Measurement Technology, vol. 2021, no. 023, pp. 044.
[9] X. Li, T. Zhang, Z. Zhang, H. Wei, and Y. Qian, "A Survey of Transformer in the Field of Computer Vision," Journal of Computer Engineering & Applications, vol. 59, no. 1, pp. 1-10, 2023.
[10] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," IEEE Transactions on Multimedia, vol. 22, no. 11, pp. 2990-3001, 2020.
[11] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," Applied Acoustics, vol. 158, p. 107020, 2020.
[12] H. Fırat, M. E. Asker, and D. Hanbay, "Hybrid 3D convolution and 2D depthwise separable convolution neural network for hyperspectral image classification," Balkan Journal of Electrical and Computer Engineering, vol. 10, no. 1, pp. 35-46, 2022.
[13] Y. Goldberg, *Neural Network Methods for Natural Language Processing*. Springer Nature, 2022.
[14] K. Kavukcuoglu, P. Sermanet, Y. L. Boureau, K. Gregor, M. Mathieu, and Y. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in Neural Information Processing Systems*, vol. 23, 2010.
[15] K. Shim, M. Lee, I. Choi, Y. Boo, and W. Sung, "SVD-softmax: Fast softmax approximation on large vocabulary neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[16] S. Savian, M. Elahi, and T. Tillo, "Optical flow estimation with deep learning, a survey on recent advances," in Deep Biometrics, 2020, pp. 257-287.

