Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/92/20241682

Deep learning approaches on computer vision

Yuandong Liu
Northeast Forestry University, Harbin, China

lyd9906@nefu.edu.cn

Abstract. In recent years, deep learning technology has made remarkable achievements in the field of computer vision, promoting the rapid development of image recognition, image generation, image understanding and other tasks. The combination of deep learning and computer vision has brought revolutionary changes to tasks such as image recognition, image generation and image understanding. This paper systematically introduces the development of deep learning and computer and the evolution of key technologies. Further, in order to give readers a deeper understanding of the field of computer vision, this article details the latest research results of some well-known scholars in computer vision, which are published in flagship conferences. The main contribution of this paper is to review the latest research results of computer vision in detail, and look forward to the future research direction and space in this field.

Keywords: Deep learning; Computer vision; Object detection; Convolutional neural networks

1. Introduction

1.1. Convolutional neural network

1.1.1. Structure and principle of CNN In recent years, deep learning technology has made remarkable achievements in the field of computer vision, promoting the rapid development of image recognition, image generation, image understanding and other tasks. Convolutional neural networks (CNN) are deep learning models specifically which designed to process 2D or 3D data such as images and videos. The structure of convolutional neural network consists of five layers: input layer, convolutional layer, pooling layer, fully connected layer and output layer. Among them, the input layer is used to receive 2D or 3D data such as images and videos. The convolution layer contains four elements: convolution kernel, activation function, step size and filling. A convolution kernel is a series of weights that are used for sliding through input data and calculating convolution. Activation functions such as Rectified Linear Unit are used to increase the nonlinearity of the model. The step size controls how far the convolution kernel moves on the input data. Padding is the addition of zeros around the input data to maintain the size of the output data. The pooling layer consists of three steps: pooling operation, step size and filling. Pooling operations, such as maximum pooling and average pooling, are used to reduce the size of the feature map and reduce parameters and computation. The step size and fill two steps are similar to the convolution layer. The fully connected layer is responsible for connecting the outputs of all feature maps to one output layer. Activation functions in this layer, such as Softmax, are used to generate classification results. The output layer is responsible for output the final prediction result, such as classification probability or regression value.

CNN extracts the local features of the image through the convolution layer and the pooling layer, and gradually constructs the global feature representation. First, the convolution layer uses the convolution

DOI: 10.54254/2755-2721/92/20241682

kernel to slide over the input data, compute the convolution result, and generate the feature map. These feature maps contain the local features of the image, such as edges, corners and so on. Secondly, the pooling layer undersamples the feature map to reduce the size of the feature map and reduce the parameters and calculation amount. At the same time, pooling operation can enhance the stability of features and make the model more robust to image rotation, scaling and other transformations. The fully connected layer then connects the outputs of all feature maps to form a global feature representation, which can learn the relationships between features to achieve image classification or regression tasks. Finally, CNN builds a multi-layer network structure by stacking multiple convolutional layers and pooling layers. Each layer of the network learns different levels of features, so as to achieve a deep understanding and recognition of images, which is the basic principle of CNN.

1.1.2. Classic CNN model Classic CNN models include AlexNet, VGG, ResNet and so on. AlexNet introduces innovations such as ReLU activation function, local response normalization layer (LRN), and GPU training, which are mainly applied to image classification and object detection. The VGG model proposes a concise and powerful network architecture that builds deeper networks by reusing the same combination of convolutional and pooling layers. VGG-16 and VGG-19 are the most famous models in this family, which are mainly used in image classification, object detection and image segmentation. ResNet solves the problem of gradient disappearance and gradient explosion in deep network training by introducing residual blocks. Residuals allow information to be passed directly from one layer to another, avoiding the loss of information during training. It is mainly used in image classification, target detection, image segmentation, attitude estimation and generation model.

1.2. Recurrent neural network

1.2.1. Structure and principle of RNN Recurrent neural network (RNN) is a special kind of neural network, which is special in that it uses neurons with self-feedback to process sequence data of any length, such as time series data, text data, etc. Recurrent neural networks are more consistent with the structure of biological neural networks than feedforward neural networks. Its structure consists of three parts: input layer, hidden layer and output layer. The input layer receives sequence data. The hidden layer, as the core part of the RNN, contains the loop structure and can process the data in the sequence. The output layer is responsible for hiding the output of the layer and generating the final prediction result.

The hidden layer in the RNN contains a loop structure that allows the network to remember previous input information, and this structure allows the RNN to handle the chronologies and dependencies in the sequence data. The hidden layer state (also known as the cell state) in the RNN is updated at each time step in the sequence to remember previous input information. Updates to the state are usually based on the current input and the state of the previous time step. At each time step, the RNN's hidden layer state is passed to the output layer, generating a prediction of the current time step. The activation function and parameters of the output layer determine the final prediction result. In addition, RNN can build a multi-layer network structure by stacking multiple hidden layers, and each layer of the network learns different levels of features, so as to achieve a deep understanding and recognition of sequence data.

1.2.2. Long short-term memory network and gated cycle unit Long short-term memory networks (LSTM) and gated cycle units (GRU) are both variants of RNN that are designed to solve the problem of disappearing gradients or exploding gradients that traditional RNN encounter when processing long sequences of data. These two models effectively deal with long-term dependencies in sequence data by introducing gating mechanisms. The core idea of LSTM is to control the flow of information by introducing three gating mechanisms. First, the forget gate determines how much information in the cell state of the previous moment should be forgotten. Second, the input gate determines how much new information should be written from the hidden state of the previous moment. Finally, the output gate determines how much information should be output from the hidden state cell state at the current

DOI: 10.54254/2755-2721/92/20241682

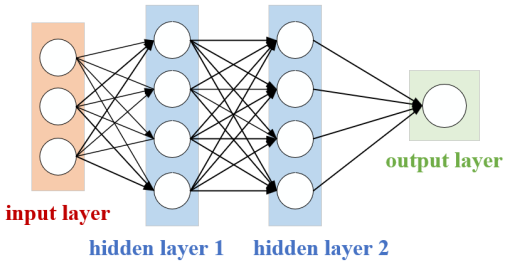

Figure 1. RNN structure diagram

time. LSTM's cell state is a long-term memory unit that allows information to persist in a sequence, thus enabling the processing of long sequence data.

The GRU combines the three gating mechanisms in the LSTM into two. The Update Gate determines how much new information should be written from the previous moment's hidden state, and the Reset Gate determines how much information should be forgotten from the previous moment's hidden state. The GRU combines the cell state and hidden state in the LSTM into a single state called "update state" and controls the flow of information through the update and reset gates.

Both LSTM and GRU are very effective at working with long sequences of data, but there are some differences between them. In terms of the number of parameters, LSTM usually has more parameters, which allows it to learn more complex sequence patterns, but also increases the complexity and computational cost of the model. In terms of computational complexity, GRU generally have lower computational complexity due to the reduction in the number of parameters, making it more practical in resource-constrained environments. As for long-term dependency processing, LSTM may have better performance when working with very long sequences because its cell state can store more long-term information. LSTM and GRU are widely used in natural language processing, speech recognition, time series prediction and other fields. They can be used for tasks such as text generation, speech synthesis, machine translation, sentiment analysis, and more. With the deepening of the research, the variants and combination forms of LSTM and GRU are also emerging, providing more choices and possibilities for the processing of sequence data.

1.3. Generate adversarial network

Generative adversarial network (GAN) consists of two main parts, generator and discriminator, which compete with each other to improve performance. The generator is used to generate realistic samples, and the discriminator is used to distinguish between generated samples and real samples. The structure of GAN can be summarized as three steps: training generator, training discriminator and iterative process. When the generator is trained, the generator becomes a data sample, the discriminator receives this generated data sample and outputs a probability value indicating the likelihood that it is real data, and then the generator adjusts its parameters based on the discriminator's feedback to produce more realistic data. The training discriminator is similar to the former, except that the discriminator also takes a sample of the real data and outputs another probability value indicating the likelihood that it is the real data. Finally, the parameters of the discriminator are adjusted based on feedback from the generator and real data samples to improve its ability to distinguish between real data and generated data. The generator and discriminator are trained alternately, and after many iterations, the generator can generate more and more realistic data, and the discriminator will become more and more discriminating.

Classical GAN models include DCGAN, WGAN, CGAN and so on. Deep Convolutional generative

DOI: 10.54254/2755-2721/92/20241682

adversarial Network (DCGAN) uses convolutional neural networks to generate and discriminate, and is suitable for image generation tasks. Wassertein Generative adversarial Network (WGAN) uses Wasserstein distance to improve the training process and improve the stability and generation quality of the generator. Conditional generation adversarial networks (CGAN) add conditional information, such as category labels or text descriptions, to generators and discriminators to generate data under specific conditions.

1.4. Large visual model and algorithm

1.4.1. Transformer Transformer model is a large language model proposed by Google in 2017[5], its core idea is the self-attention mechanism. The Transformer model is generally composed of an encoder and a decoder. Encoders are typically stacked with multiple Transformer layers and are responsible for encoding input sequences into presentation vectors. Each Transformer layer consists primarily of multi-head self-attention mechanisms and feedforward neural networks. The decoder is also stacked with multiple Transformer layers to generate the output sequence. Each Transformer layer contains a self-attention mechanism, encoder-decoder attention mechanism, and feedforward neural network. The self-attention mechanism works by calculating the similarities between each element in the sequence and the others, and using those similarities to update the representation of each element. Transformer's self-attention mechanism effectively captures long-distance dependencies between elements in a sequence, which allows Transformer models to better understand sequence data. In addition, the Transformer model is flexible, easy to understand and interpretable, and can be adapted to different tasks to efficiently process long series of data.

The Transformer model has achieved great success in natural language processing and is widely used in machine translation, text summarization, text generation, question answering systems, speech recognition, computer vision and video understanding. Multi-head attention mechanism and feedforward neural network are the key networks for image feature extraction. Transformer model achieves better results than traditional convolutional neural network in image recognition, image segmentation, object detection and other tasks.

1.4.2. Vision Transformer Vision Transformer (ViT) is a model that has made a breakthrough in the field of computer vision in recent years. It introduces Transformer model into image processing tasks, breaking the dominance of traditional convolutional neural networks in vision tasks. The structure of Vision Transformer is mainly composed of image embedding layer, Transformer encoder and classification first three parts. First, the image embedding layer divides the image into small pieces and converts them into serial data. The image features are then extracted using Transformer encoder. The Transformer encoder consists of multiple self-attention mechanisms and feedforward neural networks. Multi-head attention mechanism and feedforward neural network are the key networks to extract image features. Each layer encoder captures the global dependencies between image blocks through multi-head self-attention mechanism, and the feedforward neural network performs nonlinear transformation of the features of each image block. For the image classification task, the classification head classifies the extracted features and outputs the category of the image.

Experiments on some large image datasets, such as ImageNet, have shown that Vision Transformer achieves performance beyond traditional convolutional neural networks in image classification tasks. Especially when the amount of data is sufficient and the computing resources are abundant, the advantages of vision Transformer in terms of global receptive field, parallel computing, scalability and transfer learning are more obvious. ViT uses a self-attention mechanism that is able to capture global information in an image, rather than just local features. This allows the ViT to better understand the overall content and structure of the image. The structure of Vits is flexible, allowing for efficient processing of large scale image data and can be adjusted for different tasks. On large data sets, ViT can further improve performance by increasing the size and computational resources of the model.

DOI: 10.54254/2755-2721/92/20241682

2. Latest work in flagship conferences CVPR

2.1. Introduction

In this section, four representative papers from CVPR conferences are selected to summarize the research achievements and innovations, which all introduce the latest research progress in the field of target detection. HIG: Hierarchical Interlacement Graph Approach to Scene Graph Generation in Video Understanding was published in 2024, This paper proposes a hierarchical interlaced graph method for video understanding, which significantly improves the effect of scene graph generation. RepViT: Revisiting Mobile CNN From ViT Perspective, published in 2024, combines the strengths of vision Transformer and mobile CNN to propose a new family of lightweight convolutional neural networks suitable for visual tasks on mobile devices. Event Stream-based Visual Object Tracking: A High-Resolution Benchmark Dataset and A Novel Baseline, published in 2024, introduces a new high-resolution baseline dataset and baseline model for event stream visual object tracking, improving tracking accuracy. PREGO: Online Mistake Detection in Procedural Egocentric Videos published in 2024, this study proposes an online error detection method for detecting errors in procedural egocentric videos, improving the accuracy of automated video analysis.

2.2. Key research achievements

The paper[1] first introduces the background and challenges of visual interaction understanding problems, and points out the limitations of existing methods, such as limited interaction types and lack of modeling of complex scenes. Then it reviews relevant studies, including data sets, benchmarks, interaction modeling methods, and analyzes the limitations of existing data sets. Second, the paper proposes a new dataset, ASPIRe, which aims to address the limitations of existing datasets in visual interaction understanding. The ASPIRe dataset contains richer types of interactions, including appearance, context, location, interactions, and relationships, covering a wider range of scenarios and Settings. Next, the paper proposes a novel visual interaction understanding model called Hierarchical Interlacement Graph (HIG). The structure and training process of HIG model include hierarchical structure, unified layer, message passing mechanism, hierarchical aggregation and interactive prediction. The HIG model utilizes a hierarchical structure to capture complex interactions in video and simplifies operations and increases efficiency by unifying the layers. The HIG model can dynamically adjust its structure and function to suit different video sequences and interaction types, demonstrating strong flexibility and adaptability. In addition, the HIG model achieved state-of-the-art performance on multiple tasks, including visual interaction understanding on the ASPIRe dataset and scene graph generation on the PSG dataset, and compared with other methods, the results show that the HIG model achieves the performance of SOTA.

In summary, The paper summarizes the advantages and application prospects of HIG model, and points out the limitations of the model. The main contribution of this paper is to propose a new framework and data set for understanding visual interactivity, which is an important advance in the field of computer vision.

This work[2] proposes a new lightweight convolutional neural network family, called RepViT, which deeply analyzes the structural connections and differences between lightweight ViT and lightweight CNN, and optimizes MobileNetV3-L from multiple granularity by drawing on the design idea of ViT, and finally obtains the RepViT model. Second, RepViT uses a MetaFormer structure similar to ViT and is made up entirely of heavily parameterized convolution, allowing it to learn both global and local features efficiently while remaining lightweight. In addition, RepViT evaluated RepViT in multiple computer vision tasks and compared it to existing lightweight ViT and CNN models, which showed significant advantages in both performance and latency.

In summary, RepViT is an efficient lightweight CNN model that outperforms existing lightweight ViT and CNN models in terms of performance and latency. RepViT's success shows that lightweight CNN has great potential in mobile device deployment and provides new ideas and directions for future lightweight model research.

DOI: 10.54254/2755-2721/92/20241682

The paper [3] proposes the hierarchical cross-modal knowledge distillation strategy for the first time. HDETrack uses teacher-student network architecture and hierarchical knowledge distillation strategy to learn knowledge from multi-modal or multi-view data and transfer it to the student network using only event data to guide students' Transformer network learning. Achieve efficient, low-latency target tracking. Secondly, this paper presents the first high-resolution event tracking benchmark dataset, which provides richer data and more comprehensive evaluation criteria for event tracking research. In addition, this paper realizes cross-modal knowledge transfer, and realizes knowledge transfer from multi-mode or multi-view to single-mode event basis tracking through hierarchical knowledge distillation strategy. Finally, HDETrack has achieved excellent tracking performance on multiple data sets, and its tracking accuracy and speed are superior to existing SOTA trackers, which proves its effectiveness and robustness.

This paper has made an important contribution to the field of event tracking, realizing the transfer learning from multimodal to single-modal event tracking network. The hierarchical knowledge distillation strategy proposed by HDETrack effectively solves the problem of sparse event data and low resolution, and EventVOT dataset provides a more comprehensive evaluation standard for event tracking research. It provides new ideas and tools for event tracking research.

Firstly, an online open set error detection model PREGO is proposed in this paper[4]. PREGO is the first model capable of online identifying process errors in egocentric videos and adapting to open set scenarios. By analyzing the video step by step and predicting the next action, it can detect errors as soon as they occur. Second, PREGO uses a One-Class Classification (OCC) framework to train only on correctly executed processes, which allows PREGO to identify various process errors without being limited to predefined error types. In addition, PREGO uses a pre-trained large language model (LLM) for zero-sample symbolic reasoning to predict the next action through context analysis. This approach abstracts the video content and allows the LLM to focus on learning flow patterns. Finally, the paper evaluated PREGO's performance on Assembly101-O and Epic-tent-O datasets and compared it to multiple baseline models, showing that PREGO performed well in the online open set process error detection task.

In summary, this study proposes an online error detection method, PREGO is an innovative and practical model for detecting errors in programmatic egocentric video, improves the accuracy of automated video analysis, and provides new ideas and methods for online open set process error detection.

Taken together, these papers demonstrate significant research advances and innovations in their respective fields, from the creation of data sets to the design of algorithms, all of which make important contributions to video understanding and the development of lightweight models.

3. Challenges and development trends of computer vision

3.1. Data set training problems

One of the challenges deep learning faces in computer vision is data set issues such as insufficient data volume, uneven data distribution, poor data quality, and high cost of data annotation. Deep learning models require a large amount of training data, and insufficient data will lead to overfitting and insufficient generalization ability of the models. Unevenly distributed data can cause models to perform poorly in specific categories. Poor data quality will cause the model to learn the wrong features and affect the performance of the model.

Deep learning also has annotation problems in computer vision, such as annotation consistency, annotation bias, and low annotation efficiency. Different annotators may have different annotations on the same image, which affects the training effect of the model. The subjective factors of the tagger may lead to the deviation of tagging and affect the generalization ability of the model. Data annotation requires manual participation, which is inefficient and difficult to meet the needs of large-scale data annotation.

To solve the problem of data set and annotation, schemes such as data enhancement, data cleaning, data resampling, semi-supervised learning, self-supervised learning, automatic annotation and weakly supervised learning can be adopted. By rotating, flipping and scaling existing data, new training data can

DOI: 10.54254/2755-2721/92/20241682

be generated to increase the amount of data and improve the diversity of data. Clean up noise and wrong data in data set to improve data quality; Using data resampling technology to balance data distribution and reduce the impact of uneven data distribution; A small amount of labeled data and a large amount of unlabeled data are used to train the model to reduce the dependence on large amounts of labeled data. Using the unsupervised information of the data itself to train the model, such as autoencoder, contrast learning, to reduce the dependence on large amounts of labeled data; Automated annotation techniques, such as image segmentation and target detection, are used to improve the efficiency and accuracy of data annotation. Use weakly supervised learning techniques, such as data alignment and data matching, to reduce the dependence on annotated information.

3.2. Model complexity and computing resources

Deep learning models have made remarkable achievements in computer vision, but they also face the problem of model complexity. Model complexity refers to the amount of computation and storage of the model, as well as the number of model parameters. The training and reasoning of deep learning models require a lot of computational resources, which limits the usefulness of the models. At the same time, the model requires a large amount of storage space, which increases the difficulty of deploying the model. Training models requires a lot of time and computational resources, which limits the speed of model development.

For the problem of model complexity and computing resources, solutions such as model compression, model acceleration, lightweight models, data synthesis. Through pruning and quantization techniques, the volume and calculation amount of the model are reduced. Using GPU, TPU and other hardware to accelerate model training and reasoning. The distributed training and parallel training model are used by multiple computers. Develop lightweight models, such as MobileNet, SqueezeNet, to reduce the size and computational effort of the models.

3.3. Model interpretability and interpretability

Interpretability and interpretability challenges include model complexity, nonlinear behavior, and data complexity. The internal structure of deep learning models is complex and it is difficult to understand their decision-making process. The nonlinear behavior of deep learning models makes it difficult to analyze the relationship between their features. Computer vision data is complex and diverse and difficult to interpret in simple language. Solutions such as attention mechanisms, gradient propagation, visualization, feature importance, interpretable AI, adversarial samples. The attention mechanism can analyze the focus area of the model in the image and reveal the decision basis of the model. Gradient propagation can analyze the decision-making process of the model in the image and reveal the sensitivity of the model to the input data. Visualization can show the decision-making process of the model in the form of images or graphs, and help users understand the decision-making basis of the model. Feature importance can analyze the influence degree of each feature in the model on the prediction result and reveal the decision basis of the model. Interpretable AI aims to develop interpretable deep learning models that increase transparency and credibility. Adversarial samples can analyze the robustness of the model to the input data and reveal the potential risks of the model. Deep learning is further advanced in attention mechanisms, visualization, feature importance, interpretable AI, and adversarial samples.

3.4. Security and privacy protection

The security and privacy issues of deep learning should not be underestimated. Security issues such as anti-sample, model theft, and data privacy breaches. Adversarial samples involve adding tiny perturbations to the raw data that cause a deep learning model to make false predictions. Adversarial samples can lead to serious consequences such as misjudgment of the autonomous driving system and errors in medical diagnosis. An attacker can use model stealing technology to obtain a trained deep learning model, so as to analyze the structure and parameters of the model and understand the decision-making process of the model. When training deep learning models, a large amount of training data is

DOI: 10.54254/2755-2721/92/20241682

required. If this data contains sensitive information, it may lead to privacy breaches. Privacy protection issues such as data collection, data annotation and model training. When collecting user data, it is necessary to ensure the legality and privacy of the data to avoid unauthorized data collection and abuse. During data annotation, it is necessary to protect the privacy of the tagger and avoid the disclosure of the tagger's identity information. In the process of model training, it is necessary to protect the privacy of training data and avoid the leakage of training data.

For the security and privacy protection of deep learning, measures can be taken in the aspects of anti-sample defense, model security, data privacy protection, privacy computing and compliance. Adversarial sample defense techniques such as adversarial training, adversarial data enhancement and input space constraint are used to improve the robustness of the model to adversarial samples. Model encryption, model obfuscation and other technologies are used to protect the structure and parameters of the model and prevent model theft. Use differential privacy, federated learning and other technologies to protect data privacy during data collection, annotation and training. Privacy computing techniques, such as state encryption and secure multi-party computing, are used to protect data privacy during model training and prediction. Comply with relevant laws, regulations and standards, such as the EU's GDPR, to ensure compliance with data processing and model application.

Deep learning will continue to evolve in the future in terms of adversarial sample defense, model security, privacy computing, and compliance. Adversarial sample defense technology will continue to develop and improve to cope with higher level adversarial sample attacks; Model security will become an important consideration in the design and application of deep learning models. Privacy computing technology will continue to improve, providing better privacy protection solutions for deep learning models; As data privacy regulations continue to improve, compliance will become an important consideration in the design and application of deep learning models.

4. Conclusion

4.1. Summary

This paper summarizes the structure and basic principles of classical deep learning models. Through in-depth study of deep learning models, we can better understand their principles and advantages in computer vision tasks and improve their application effects in various fields. At the same time, by summarizing the latest research achievements and innovations in CVPR conference, this paper expounds the importance and influence of deep learning technology in computer vision tasks such as image recognition, object detection, image segmentation, image generation and image description. In various tasks of computer vision, deep learning models show strong performance and application potential. Deep learning models, especially CNN and Transformer, have made remarkable progress in the field of computer vision. These models can automatically extract image features and process nonlinear data, which provides a strong support for the development of computer vision technology.

However, the application of deep learning in computer vision also faces some challenges, such as computational resource requirements, model complexity and interpretability. To address these challenges, researchers have proposed a number of improved deep learning models and algorithms, such as model compression, model acceleration, and interpretable AI. These improved models and algorithms help to improve the application of deep learning in computer vision and promote the development of the technology. With the deepening of research, the application of deep learning in computer vision will be more extensive. In the future, the development trend of deep learning technology in the field of computer vision will include multi-modal learning, small sample learning, security and privacy protection. These trends will contribute to the continued development of deep learning in the field of computer vision, bringing innovation and progress to more fields.

In summary, this paper summarizes the application, importance, challenges, solutions and development trends of deep learning technology in computer vision tasks. As technology continues to evolve and algorithms continue to be optimized, deep learning will continue to play an important role in the field of computer vision, bringing more convenience and value to humans.

DOI: 10.54254/2755-2721/92/20241682

4.2. Future research direction

Deep learning has made remarkable progress in the field of computer vision, but there are still many challenges and research directions. The future of deep learning will continue to develop in the direction of small sample learning, model interpretability, model security, cross-modal learning, computational efficiency, hardware and software co-design, deep learning exploration in new application fields, artificial intelligence ethics and social impact. Improve the performance of deep learning models on small sample data, which is of great significance for the rapid adaptation of new tasks and domain adaptation. Develop deep learning models that can explain their decision-making processes to improve the transparency and credibility of the models. Investigate how to improve the robustness of deep learning models against adversarial samples and how to protect models from attacks. Combine multiple modes of data (such as images, text, audio, etc.) to improve model understanding and reasoning. Develop deep learning models that can handle multiple related tasks simultaneously to improve the generalization ability and efficiency of the models. Investigate how to train deep learning models with a small amount of labeled data and a large amount of unlabeled data to reduce dependence on labeled data. Develop more efficient deep learning models and algorithms to reduce computing resources and energy consumption. Explore the co-design of deep learning model and hardware (such as GPU, TPU, FPGA, etc.) to improve the computational efficiency and practicability of the model. Explore the potential of deep learning in biomedical image analysis, natural language processing, robot vision and other novel applications. Investigate the ethical and societal implications of the application of deep learning techniques in computer vision, and how to ensure that the development of the technology meets ethical standards and societal needs. With the continuous deepening of research, the future research direction of deep learning in the field of computer vision will be more extensive and in-depth, and then promote the continuous development of computer vision technology.

References

[1] Trong-Thuan Nguyen, Pha Nguyen, Khoa Luu, HIG: Hierarchical Interlacement Graph Approach to Scene Graph Generation in Video Understanding, CVPR 2024.

[2] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, Guiguang Ding, RepViT: Revisiting Mobile CNN From ViT Perspective, CVPR 2024 Camera-ready Version.

[3] Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, Jin Tang, Event Stream-based Visual Object Tracking: A High-Resolution Benchmark Dataset and A Novel Baseline, CVPR 2024.

[4] Alessandro Flaborea, Guido Maria D' Amely di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Galasso, PREGO: Online Mistake Detection in Procedural Egocentric Videos, CVPR 2024.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need[J].arXiv, 2017.