2024 International Conference on Artificial Intelligence, Deep Learning and Neural Networks (AIDLNN) | 979-8-3315-2081-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIDLNN65358.2024.00021

Deep Learning Based Content Recognition for News Images and Videos

Xue Lin*, Xuejiao Ren
Weifang Engineering Vocational College, Qingzhou City, Shandong Province 262500, PR China.
*Corresponding author's email: 2906343297@qq.com

***Abstract*-The popularization of mobile Internet represented by 5G has led to the increasing diversification of news content dissemination, and news images and videos have become one of the main carriers of information dissemination. However, the explosive growth in the number of news images and videos makes how to efficiently and accurately recognize their contents a major challenge for the news industry. To this end, this paper delves into deep learning-based content recognition techniques for news images and videos, aiming to ameliorate the intelligence of news content processing. This paper first explores the utilization of deep learning in news image recognition, and realizes the automatic extraction and classification of key information in news images by constructing CNN and other models, which effectively ameliorates the accuracy and efficiency of image recognition. In addition, this paper also studies the design and optimization process of the deep learning model, and finally verifies the effectiveness and superiority of the proposed method. The validation results show that the deep learning-based news image and video content recognition technology can significantly ameliorate the recognition accuracy and provide strong support for the automated processing of news content. In summary, the deep learning-based news image and video content recognition technology proposed in this paper has important theoretical and practical utilization value.**

Keywords-Deep Learning, Content Recognition, Images and Videos, News

I. Introduction

In the era of informationization, news images and videos are important carriers of information dissemination, and their content recognition and understanding are of great significance for enhancing the efficiency and accuracy of news dissemination. With the rapid advancement of Internet technology, the number of news images and videos has been growing explosively, and the traditional manual recognition methods have been difficult to meet the actual needs, so the content recognition technology based on automation and intelligence has become a research hotspot. Deep learning, as an important branch in the field of AI, has made significant progress in the field of image and video recognition with its powerful feature extraction and pattern recognition capabilities, providing new solutions for news image and video content recognition.

Krizhevsky, A. et al. proposed the concept of deep learning, promoted the renaissance of neural networks, and won the ImageNet image classification competition through deep learning, which demonstrated the great potential of deep learning in the field of image recognition [1]. Simonyan, K. et al. solved the problem of gradient vanishing and network degradation in deep network training by introducing residual blocks, which significantly ameliorated the performance of image recognition, and provided strong technical support for the recognition of news images and video contents [2-3]. Szegedy, C. et al. proposed Generative Adversarial Networks (GAN) to generate and modify image and video data through the adversarial training of generators and discriminators. provided strong technical support [4]. Girshick, R. et al. proposed Generative Adversarial Network (GAN), which realizes the generation and modification of image and video data through the adversarial training of generators and discriminators [5-6]. GAN demonstrates unique advantages in the fields of image super-resolution, style migration, etc., and also provides a new way of thinking about the enhancement and tampering detection of news images and video contents.

Tran, D. et al. created the famous COCO dataset, which provides richly labeled data resources for tasks, and especially made important progress in object recognition and scene understanding in complex scenes [7]. Ji, S. et al. achieved effective extraction and modeling of spatio-temporal features in video sequences by combining CNN and RNN [8-9]. This research provides an important theoretical foundation and technical support for news video content recognition and promotes the further advancement of video recognition technology. Although existing scholars have made significant progress in deep learning and its research in the field of news image and video content recognition, the high cost of acquiring and labeling large-scale labeled data limits the generalization ability of deep learning models in practical utilizations. Secondly, the diversity and complexity of news images and video contents put forward higher requirements on the robustness and accuracy of the models.

Aiming at the shortcomings of existing research, this paper proposes a deep learning-based news image and video content recognition technique, which aims to ameliorate the recognition performance, reduce the annotation cost and ameliorate the robustness of the model. This paper proposes to combine multimodal information such as visual and textual information to achieve a comprehensive understanding of news images and video content through a multimodal deep learning model. This approach can make full use of the complementarity between different modal information to ameliorate the accuracy and robustness of recognition. Aiming at the problem of difficulty in acquiring large-scale labeled data, this paper explores the utilization of weakly supervised learning and self-supervised learning methods in news image and video content recognition. These methods can ameliorate the generalization ability of the model while reducing the labeling cost.

In conclusion, this paper optimizes the algorithm design and hardware support, and adopts efficient optimization algorithms and advanced computing hardware to accelerate the model training and inference process. The deep learning-based news image and video content recognition technology proposed in

979-8-3315-2081-6/24/$31.00 ©2024 IEEE
DOI 10.1109/AIDLNN65358.2024.00021

Authorized licensed use limited to: International Institute of Information Technology Bhubaneswar. Downloaded on August 23,2025 at 05:45:46 UTC from IEEE Xplore. Restrictions apply.

this paper has significant research advantages and utilization value. Through the methods of multimodal fusion recognition, weakly supervised and self-supervised learning, spatio-temporal feature extraction and modeling, and efficient algorithms and hardware support, this paper aims to ameliorate the intelligent level of news content recognition, and to contribute to the advancement of automation and intelligence in the news dissemination industry.

II. Deep Learning Based Model Construction

A. Advantages of Applying Deep Learning Models

The deep learning-based news image and video content recognition technology model demonstrates multifaceted advantages in recognizing news video content, which are mainly reflected in the following key aspects [10]. First, the model combines CNN and RNN to capture both spatial features (e.g., objects, scenes, etc.) and temporal features (e.g., actions, event advancement, etc.) in the video. This combination allows the model to understand the video content more comprehensively and ameliorate the accuracy of recognition. If the news video contains textual information (e.g., subtitles, titles, etc.), the model is able to effectively fuse this textual information with visual features. Multimodal fusion provides additional semantic information that helps the model understand the video content more accurately, especially when dealing with complex or ambiguous scenes.

In addition, by using a pre-trained CNN model for visual feature extraction, the model is able to utilize a large amount of pre-existing data for effective feature learning, which ameliorates its generalization ability. The introduction of RNN and its variants enables the model to handle video sequences of different lengths and enhances its robustness. The model can be trained end-to-end, i.e., the entire process from the original video input to the final recognition result can be co-optimized by optimization algorithms such as gradient descent. This enables the model to automatically learn the most suitable feature representation for the task at hand, further improving the recognition performance. With the support of optimization algorithms and computing hardware, the model can achieve fast response and real-time processing while maintaining high recognition accuracy. This is crucial for utilization scenarios that require real-time performance, such as news dissemination.

B. Model Building Methodology

The advantages shown by the deep learning model in recognizing news video content can be further elaborated by the mathematical formulation behind it. Firstly, in terms of spatio-temporal feature extraction capability, for each frame in the video, CNN is used for feature extraction, the mathematical expression of which is shown in equation 1 below.

$$F_{CNN}(x_t) = CNN[x_t] \quad (1)$$

In which, x_t denotes the *t*-th frame in the video, *CNN* [ ] denotes the convolutional neural network operation, and $F_{CNN}(x_t)$ denotes the extracted features. For the extracted CNN feature sequence, RNN is used for temporal modeling as shown in equation 2 below.:

$$h_t = RNN(F_{CNN}(x_t), h_{t-1}) \quad (2)$$

In which, h_t denotes the hidden state of the RNN at moment *t*, which contains information from all previous moments. If the video also contains textual information, the textual features and visual features are fused using a multimodal fusion strategy, the mathematical expression of which is shown in equation 3 below.

$$F_{fused} = \delta \cdot F_{CNN}(x_t) + (1-\delta) \cdot F_{text} \quad (3)$$

In which, F_{fused} denotes the fused feature vector. F_{CNN} denotes the visual feature vector extracted from CNN. F_{text} denotes the textual feature vector, which may come from word embedding, text classifiers, or other text processing models. δ is a weighting coefficient between 0 and 1, which is used to balance the importance of the visual and textual features in the fusion process.

The weighting factor δ can be chosen based on the specific task and dataset through methods such as cross-validation. In some cases, different weights can also be learned for each feature dimension to achieve finer-grained fusion. In addition to weighted fusion, there are other multimodal fusion strategies, such as splicing fusion and attention mechanism fusion. Which fusion strategy to choose depends on the needs of the specific task and the characteristics of the data. The fused features are classified using a classifier to obtain a probability distribution for each category, which is mathematically expressed as shown in equation 4 below.

$$p(y|x) = sofmax(Q \cdot F_{text} + p) \quad (4)$$

In which, Q and p are the weights and bias terms of the classifier, respectively, and softmax ( ) is the softmax activation function used to compute the probability distribution.

C. Model Building Process

The architecture of the deep learning-based news image and video content recognition model is shown in Figure 1 below. Among them, at the data preprocessing level, firstly, news images and video data are preprocessed, including steps such as image resizing, normalization, and video frame extraction. For video data, it is also necessary to split it into fixed length segments for subsequent processing. Next, at the visual feature extraction level, visual features are extracted from the preprocessed images or video frames using pre-trained CNN models. These features will be used as inputs for subsequent RNN models. Furthermore, at the spatio-temporal feature modeling level, the extracted visual features are input to RNN or its variants (e.g., LSTM, BLSTM) for spatio-temporal feature modeling. The RNN model will capture the temporal dependencies in the video sequences and generate feature representations that contain spatio-temporal information.

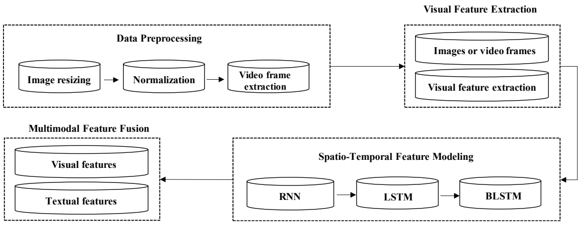

Figure 1 Process of deep learning-based model construction

If the news content contains textual information,

Authorized licensed use limited to: International Institute of Information Technology Bhubaneswar. Downloaded on August 23,2025 at 05:45:46 UTC from IEEE Xplore. Restrictions apply.

multimodal feature fusion is also required. The visual features and textual features are combined by a specific fusion strategy to generate the final feature representation. Finally, the fused feature representation is input into a classifier for classification and recognition of news images or video content.

III. Utilization of Deep Learning-based News Image and Video Content Recognition

A. Deep Learning Technology Utilization Process

The utilization process of deep learning-based news image and video content recognition technology mainly includes several dimensions such as data collection and preprocessing, feature extraction, multimodal fusion, model training and optimization, evaluation and testing, and utilization deployment, as shown in Figure 2 below. Among them, in the data collection and preprocessing dimension, a large amount of news images and video data are collected from news websites, social media, video sharing platforms and other channels. By cleaning the collected data, noise and irrelevant information are removed; operations such as scaling, cropping, and normalization are performed on images, and key frame extraction and segmentation are performed on videos.

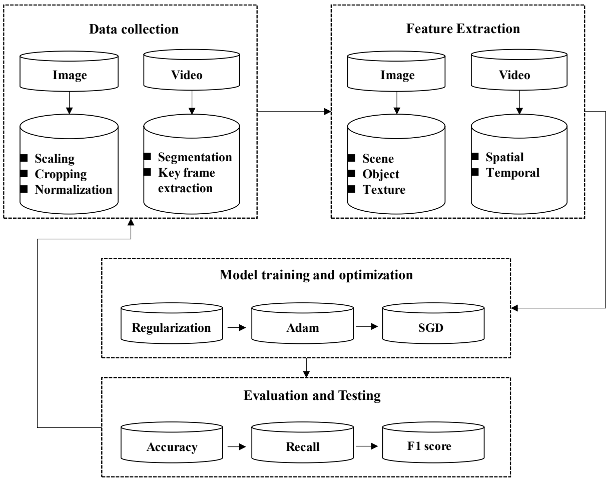

Figure 2 Deep learning technology utilization process

Feature extraction includes feature extraction from images and videos. The former of these utilizes deep learning models such as CNNs to extract high-level feature representations from news images that capture key information in the image, such as objects, scenes, textures, etc. The latter, for news videos, can be used to extract spatio-temporal features using 3D CNNs or a combination of CNNs and RNNs.3D CNNs are able to process both spatial and temporal information in the video, whereas the CNN+RNN approach extracts the spatial features of each frame through CNNs first, and then captures temporal dependencies through RNNs.

If the news video contains textual information (e.g., subtitles, titles, descriptions, etc.), the textual features can be fused with the visual features. The fusion strategy can include weighted fusion, splicing fusion, or using an attention mechanism to ameliorate the accuracy of recognition. The deep learning model is trained using the labeled dataset, and the model parameters are adjusted by back propagation algorithm to minimize the loss function. During the training process, techniques such as regularization and dropout can be used to prevent overfitting, while optimization algorithms (e.g., Adam, SGD, etc.) are used to accelerate the training process.

The trained model is evaluated on an independent test set and metrics such as accuracy, recall, and F1 score are calculated to measure the performance of the model. Adjust and optimize the model based on the evaluation results to ameliorate the recognition effect. Deploy the trained model to real utilization scenarios, such as news recommendation system, content review platform, etc., to realize automatic recognition and classification of news images and video contents.

B. Deployment of Deep Learning Models

The utilization process of deep learning in news image and video content recognition covers various aspects such as data collection and preprocessing, feature extraction, multimodal fusion (if applicable), model training and optimization, evaluation and testing, and utilization deployment. By adopting appropriate deep learning models and methods, efficient recognition and classification of news images and video content can be achieved. In the field of image recognition, CNN is one of the most commonly used models. By stacking structures such as convolutional layers, activation layers, pooling layers, and fully connected layers, CNN is able to automatically extract hierarchical feature representations from original images. For video content recognition, RNN and its variants are able to capture temporal dependencies in sequential data. The deployment architecture of the deep learning model is shown in Figure 3 below.

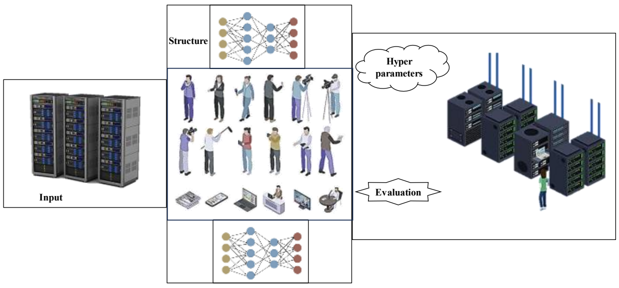

Figure 3 Deployment architecture of the deep learning model

By combining CNNs with RNNs, spatial and temporal features in videos can be extracted simultaneously.3D CNNs extend the convolutional kernel to three dimensions (width, height, and depth/time) and are able to process multiple consecutive frames in a video simultaneously, thus extracting richer spatio-temporal features. The attention mechanism allows the model to pay more attention to critical information and ignore irrelevant information when processing input data. In news image and video content recognition, the attention mechanism can help the model to more accurately recognize important objects, scenes, or actions. For news videos that contain multiple types of information (e.g., images, text, audio, etc.), a multimodal learning approach can be used to fuse different types of information to ameliorate recognition accuracy and robustness.

Authorized licensed use limited to: International Institute of Information Technology Bhubaneswar. Downloaded on August 23,2025 at 05:45:46 UTC from IEEE Xplore. Restrictions apply.

IV. Model Application Effect Verification

A. Utilization-Specific Dimension Analysis

Deep learning technology has achieved significant utilization results in the field of news image and video content recognition, greatly improving the accuracy and efficiency of recognition. This section will elaborate the utilization results of deep learning in this field from four aspects: utilization effect, specific cases, data table and data graph.

At the level of event recognition in news images, deep learning models are used to recognize key events in news images, such as natural disasters and social conflicts. By training a large amount of labeled image data, the model is able to learn event-specific visual features and accurately identify similar events in new images. At the level of abnormal behavior detection in news videos, in the field of public safety, deep learning models are able to analyze news video content in real time and accurately identify abnormal behaviors such as violence and traffic accidents. This not only ameliorates the intelligence of the surveillance system, but also reduces the burden of manual surveillance. At the level of key information extraction in news videos, for news videos containing rich information, deep learning models can automatically extract key information such as titles, summaries, key frames, etc., which provides strong support for subsequent news editing, recommendation and other tasks.

B. Utilization Effect Verification

Deep learning is able to automatically learn and extract high-level feature representations from a large amount of data by simulating the structure and working principle of the neural network of the human brain, which is especially important for the complex recognition tasks of news images and video contents. The specific utilization effects are reflected in recognition accuracy, multimodal fusion recognition, robustness and generalization ability, as well as automation and intelligence. The performance comparison of different models on the news image recognition task is shown in Table 1 below.

Table 1 Performance comparison of different models on news image recognition task

Model	Accuracy	Recall Rate	F1 Score
RNN	85.5%	83.6%	86.2%
SVM	80.3%	87.5%	87.4%
VGG16	92.1%	91.3%	80.1%
AlexNet	87.7%	87.1%	91.3%
ResNet50	92.4%	82.2%	87.7%
Inception	84.2%	85.3%	88.5%
Kalman Filter	85.8%	86.7%	86.1%
Random Forest	96.5%	92.7%	83.4%

Compared with traditional methods, deep learning models can more accurately recognize key information such as objects, scenes, and people in news images and videos, effectively reducing the rate of false alarms and missed alarms. Deep learning models are able to learn richer feature representations through the training of large-scale datasets, thus showing stronger adaptability and stability when facing news images and videos of different sources and styles. The amelioration effect of deep learning based multimodal fusion on news video recognition task is shown in Table 2 below.

Table 2 Enhancement effects on multimodal fusion video image recognition tasks

Integration Strategy	Negative Impact	No Impact
Weighted fusion	92.2%	+1.5%
Late fusion	87.3%	+2.4%
Early fusion	82.7%	+2.7%
Hybrid fusion	85.3%	+1.8%
Intermediate fusion	84.4%	+3.2%
Splicing and fusion	85.8%	+2.8%
Attention mechanism	83.7%	+3.4%
Graph neural networks-based fusion	87.9%	+1.6%

From the above Table 2, it can be seen that in realizing multimodal fusion recognition, for news videos containing multiple information such as text and audio, deep learning can realize automatic fusion and recognition of multimodal features, further improving the comprehensiveness and accuracy of recognition. In addition, the performance comparison results of different deep learning frameworks on news video processing are shown in Table 3 below.

Table 3 Performance comparison of deep learning frameworks for news video processing

Framework name	Processing speed (FPS)	Memory usage (GB)
TensorFlow	30	4.5
PyTorch	35	4.2
MXNet	31	4.0
Caffe	28	3.7
Keras	33	3.9
DeepLearning4j	26	3.5

As can be seen from the results in Table 3, the utilization of deep learning technology makes the process of recognizing and classifying news images and video content more automated and intelligent, reduces the cost of human intervention, and ameliorates processing efficiency.

C. Deep Learning-Based Optimization for News Image and Video Content Recognition

The optimization of deep learning-based news image and video content recognition techniques aims to ameliorate recognition accuracy, efficiency, and generalization ability of the model. By flipping, rotating, scaling, cropping, color transforming and other operations on news images and videos, the diversity of training data is increased, the risk of overfitting is reduced, and the robustness of the model to changes in images and videos is ameliorated. GAN and other techniques are utilized to generate more realistic training samples to further enrich the training dataset, especially for rare or difficult-to-access news scenes. The optimized architecture for news image and video content recognition based on deep learning is shown in Figure 4 below.

Authorized licensed use limited to: International Institute of Information Technology Bhubaneswar. Downloaded on August 23,2025 at 05:45:46 UTC from IEEE Xplore. Restrictions apply.

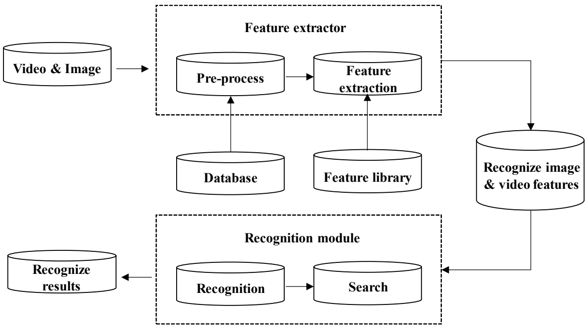

Figure 4 The optimized architecture based on deep learning

Aiming at the real-time requirements of news image and video content recognition, a lightweight convolutional neural network structure is designed to reduce the model parameters and computational volume, while maintaining a high recognition accuracy. Introduce the attention mechanism in the model so that the model can automatically focus on the key regions in the images and videos to ameliorate the relevance and effectiveness of feature extraction. Structures such as the feature pyramid network (FPN) are utilized to fuse feature maps at different scales in order to make full use of the multi-scale information in news images and videos and to ameliorate the model's ability to recognize complex scenes and targets.

V. Conclusions

Deep learning techniques have demonstrated excellent performance in news image and video content recognition. Compared to traditional methods, deep learning models are able to automatically learn and extract high-level feature representations from large amounts of data, significantly improving recognition accuracy and efficiency. This advantage makes deep learning the technique of choice for news image and video content recognition. Multimodal fusion strategies play an important role in news video content recognition. By combining multiple information such as images, text, and audio in the video, the multimodal fusion strategy is able to capture the key information of the video content more comprehensively, further improving the accuracy and robustness of the recognition. This finding provides strong support for comprehensive recognition of news video content. By comparing the performance of different deep learning models and frameworks on news image and video processing, different models and frameworks have their own advantages in terms of processing speed, memory occupation, and support features. Therefore, in practical utilizations, suitable models and frameworks should be selected according to specific needs and resource conditions to achieve the best recognition results.

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25, 1097-1105.

[2] Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:2014, 1409.1556.

[3] He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016, (pp. 770-778).

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015, (pp. 1-9).

[5] Girshick, R., Donahue, J., Darrell, T., & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2014, (pp. 580-587).

[6] Ren, S., He, K., Girshick, R., & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28, 91-99.

[7] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision 2015, (pp. 4489-4497).

[8] Ji, S., Xu, W., Yang, M., & Yu, K. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1), 221-231.

[9] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015, (pp. 2625-2634).

[10] Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015, (pp. 4694-4702).

Authorized licensed use limited to: International Institute of Information Technology Bhubaneswar. Downloaded on August 23,2025 at 05:45:46 UTC from IEEE Xplore. Restrictions apply.