JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Top-Heavy CapsNets Based on Spatiotemporal Non-Local for Action Recognition

Manh-Hung Ha *

Faculty of Applied Sciences, International School, Vietnam National University, Hanoi 100000, Vietnam; e-mail: hunghm@vnu.edu.vn
* Corresponding Author : Manh-Hung Ha

Abstract: To effectively comprehend human actions, we have developed a Deep Neural Network (DNN) that utilizes inner spatiotemporal non-locality to capture meaningful semantic context for efficient action identification. This work introduces the Top-Heavy CapsNet as a novel approach for video analysis, incorporating a 3D Convolutional Neural Network (3DCNN) to apply the thematic actions of local classifiers for effective classification based on motion from the spatiotemporal context in videos. This DNN comprises multiple layers, including 3D Convolutional Neural Network (3DCNN), Spatial Depth-Based Non-Local (SBN) layer, and Deep Capsule (DCapsNet). Firstly, the 3DCNN extracts structured and semantic information from RGB and optical flow streams. Secondly, the SBN layer processes feature blocks with spatial depth to emphasize visually advantageous cues, potentially aiding in action differentiation. Finally, DCapsNet is more effective in exploiting vectorized prominent features to represent objects and various action features for the ultimate label determination. Experimental results demonstrate that the proposed DNN achieves an average accuracy of 97.6%, surpassing conventional DNNs on the traffic police dataset. Furthermore, the proposed DNN attains average accuracies of 98.3% and 80.7% on the UCF101 and HMDB51 datasets, respectively. This underscores the applicability of the proposed DNN for effectively recognizing diverse actions performed by subjects in videos.

Keywords: Action recognition; Attention mechanism; Capsule network; Deep neural network; Spatiotemporal.

1. Introduction

Deep Neural Networks (DNN) have demonstrated remarkable success in addressing visual recognition problems that are extensively handled to extract visual features and have been widely and extremely good outcomes in video understanding. Nevertheless, the powerful class of Convolution Neural Network (CNN) model research on video recognition has been adopted in local operations related to spatial representation while ignoring the in-formation intensive of complex temporal variations, which rely on the pattern of abundant spatial and temporal feature [1]–[4]. As 2D CNNs only capture individual appearance features for each frame without considering motion information in video sequences, mul-ti-stream CNNs[1]–[3], [5] have been introduced. These networks incorporate stacked optical flow (OF), RGB, depth maps, etc., as additional inputs to enhance short-term modeling. However, CNN still has several limitations. It is tough to gather diverse training data, the limited equivalence, and the inability to maintain spatial hierarchies' features, effectively concentrate on, and efficiently contribute useful clues to the learning processes. Typically, previous CNN-based methods of video action recognition involved two key steps: creating frame-level action proposals and connecting ideas across frames. Moreover, most of these algorithms use a two-stream CNN architecture to separate spatial and temporal information. In contrast to 2D CNNs, 3D CNNs[6], [7] utilize 3D spatiotemporal kernels to collect both spatial and temporal information simultaneously, making them more appropriate for video analysis. They are widely employed in action recognition as well as various improved methods have emerged from 3D CNNs[6], [8].

Received: April, 29th 2024
Revised: May, 13th 2024
Accepted: May, 24th 20204
Published: May, 25th 2024

Recently, attention mechanisms have been incorporated into models to account for global dependencies [6], [9]. Attention mechanisms are crucial for capturing long-range dependencies across the entire image and have been used to enhance the performance of CNNs in tasks such as image classification and scene segmentation [10], [11]. In these works, attention mechanisms are applied both spatially and channel-wise. Spatial attention focuses on spatial connections between geographical objects, while channel attention enhances certain channels' importance and diminishes others' importance. The Squeeze-and-Excitation module [12] serves as a channel-specific attention block, improving CNN performance with minimal computational cost. On the other hand, the non-local block [13]–[15] leverages spatial and temporal information to capture dependencies between features across frames, akin to an attention-based approach. However, previous research has indicated that attention mechanisms are primarily used to handle two-dimensional information related to spatial attention mechanisms. The attention module computes the spatial correlation matrix between any two locations in the input feature maps. Each position is computed and updated using the weighted sum of the previous positions. To enhance the 3D spatial-temporal modeling in videos, we have de-vised a 3D attention model that combines spatial and depth attention modules to capture feature-level correlations in three dimensions. The proposed SBN neural network enables convolutional layers to capture long-range relationships in 3D space, enhancing the recognition performance of actions.

Capsule Network (CapsNet), a group of capsules instead of a neuron to encode the part-whole relationship, was first proposed as a new architecture by GE Hinton et al. [16]. As the original CapsNet relies solely on shallow CNNs to retain spatial information, the introduction of deep CapsNet [17]–[19]aimed to address the absence of semantic information. Although research on capsules is intuitive and has many fundamental advantages, a notable issue in performance arises as learning becomes more challenging with deeper CapsNets. This difficulty may stem from issues such as gradient explosion or an elongated chain of dynamic routing. Deep CapsNet may require higher dimensions of capsules and increase the number of trainable parameters when classifying complex datasets [17]. Therefore, in this study, we proposed top-heavy CapsNets in which Deep CapsNet is the tail of architecture.

The motivation from the involved series of technical are considered to address these gaps. Firstly, the 3D CNN [3] was proposed to capture spatial and temporal information from appearance and motion features, which are more appropriate for video recognition. Secondly, spatial and temporal non-local attention use the correlation transposed space feature to mimic the human perception that selectively concentrates on significant regions within the visual space to gather information for better understanding.

2. Related Works

Several research efforts have been successful in creating spatiotemporal features for action classification, including the use of two-stream networks combining RGB with Optical Flow [3], [5], [20]. Moreover, two streams based on 3D CNNs trained on large-scale, high-quality datasets such as ImgNet, Sports 1M, and Kinetics allow the training of deeper 3D CNN models and achieve high performance [3], [20], [21]. I3D holds one of the best architectures for action recognition on large-scale Kinetics datasets. By adopting various I3D techniques based on strategies like pose motion, motion augmentation, and IDT for visual illusion [8], [20], [21], the results have been significantly improved. This motivation inspired us to leverage a three-stream CNN architecture for multi-stream learning.

Inspired by machine translation and object detection, Xu et al. [22]introduced an automatic attention mechanism to learn images' content. This involved using an attention mechanism in the DNN to emphasize the meaningful part of an image. In [1], an attention mechanism for person action detection was proposed, focusing on modeling the surrounding context of actors. Attention functions for relating different positions in a sentence were employed in [6], [11], emphasizing meaningful regions. Additionally, self-attention modules were introduced in [1], [2], utilizing object relations to detect them and effectively enhance results in image generation. As a result of these positional relations-based attention models, we developed a spatio-depth attention model that scales the feature map at various locations according to the channel feature for action recognition.

To enhance performance, common structures are employed to integrate convolutional networks and CapsNet [3], [17]–[19]. The primary capsule and convolutional layers are

combined to improve efficiency using self-attention features. For image classification, [19] stacks multiple residual blocks and utilizes capsule layers for shared transformation matrices. Additionally, a deeper design can be employed to enhance the performance of the capsule network when dealing with complex data. Our work leverages hierarchical relationships within the CapsNet level to make predictions more effectively in the final part of the efficient architecture.

Our contribution to this paper is as follows:

- We proposed developing a new architecture for video activity recognition that is able to be composed, both of which exploit rich spatial and temporal feature abstraction through an attention mechanism. This proposed top-heavy CapsNets, in which Deep CapsNet is the tail of architecture present, enhances performance and enables more accessible learning.
- We conducted the effectiveness approach through extensive experiments to obtain a good architecture in the case with and without capsnet, and achieve potential outcomes on the UCF 101 dataset.

3. Proposed Model DNN for Action Recognition

In this study, we present the proposed DNN for action recognition. Our DNN framework consists of the I3D backbone, SBN creation layers, and CapsNet, as illustrated in Figure 1 (b). During preprocessing, the RGB stream is evenly divided into one or multiple temporal frame segments T, with a 50% overlap ratio. The optical flow channel frames are generated from the RGB frames using the well-known regularization and robust L1 norm (TVL1) [3]. Each pair of X-axis and Y-axis optical flow channels has dimensions (224×224×2). The linear transformation further normalizes each stacked component in these two optical flow channels to a value between 0 and 255.

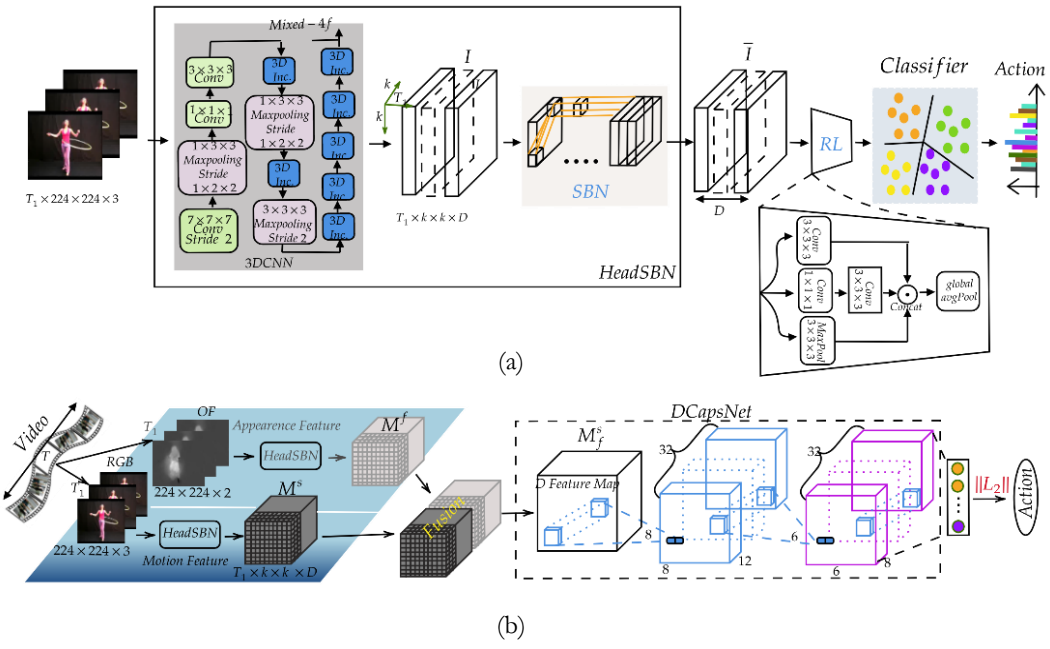

Figure 1. Block diagram of the proposed DNN. Our architecture is built upon a 3D CNN network up to the 'Mixed-4f' block to generate spatial and temporal feature maps; (**a**) Appearance stream structure, where we apply the Spatial Bottleneck (SBN) layer, Reduction Layer (RL), and classification layer for the final prediction; (**b**) Motion stream structure, where we employ feature extraction through HeadSBN with both appearance and motion streams, and then CapsNet trains deep correlations as a tail layer to enhance meaningful action classification.

3.1. Proposed Model DNN-1

The proposed DNN comprises the SBN layer following the 3D CNN (referred to as headSBN), the Reduction Layer (RL), and the classifier, as depicted in Figure 1 (a). The headSBN layer employs the 3D CNN with an Inception 3D (I3D) structure from the input layer to 'Mixed-4f' to create feature blocks through transfer learning. Features are commonly

extracted from the 'Mixed-4f' module and passed to the SBN layer to generate corresponding spatial and motion feature maps. An RL frame is proposed to combine the receptive fields of pooling and convolutional layers to reduce dimensionality. 1×1×1 3D convolution filters are utilized to reduce the number of input channels before larger 3×3×3 convolutions, making it computationally less expensive than alternatives while preserving correlation information. The final layer determination is achieved through the fully connected (FC) layer inference classifier. Video clips are segmented into multiple segments consisting of 10 RGB frames as input.

3.1.1. Non-local Attention (NA)

To assist in identifying crucial elements of an activity, a local neural attention block is depicted in [15] to highlight discernible components of activity and leverage existing spatial attention units to establish spatial relationships in the feature map. The NA module generates object maps U and Q by passing input from the 3D spatial object map I with dimensions T_1×k×k×D through two distinct 3D convolution layers with kernels of size 1×1×1. Subsequently, utilizing matrix multiplication, U and Q determine attention weights, where D is the number of feature channels, and T_1 represents the spa-tial-temporal volume. k×k denotes the number of spatial feature maps. The data is passed through another 3D convolution layer with a kernel size of 1×1×1 to construct the feature map. By multiplying K and attention weights, we obtain a new feature map. Additionally, a scaling parameter and the feature I are introduced to ensure the consistency of the spatial attention output.

The Attention Layer combines individual object feature maps, synthesizing the data to generate attention weights for integrating feature maps and producing the final output. The physical significance is that traces of certain activities are contingent on a specific theme rather than the entire context. Thus, individual maps with emphasized logical region relationships may benefit activity discrimination. Different thematic regions of features are carefully combined using attention mechanisms, ensuring that broad contextual associations are adequately captured during the classification process.

3.1.2. The proposed Spatial-Depth Based Non-Local (SBN)

In Figure 2, 3D spatial and depth non-local modules are included in our 3D SBN module. According to our definition of feature maps, I, as having the dimension $T_1 \times k \times k \times D$, where dimensions T_1, D denoted depth and channel number, respectively, $k \times k$ denotes the height and width of a spatial feature map, as seen in Fig. 2. This is done by first applying a 1×1×1 kernel on the input feature map, I, and creating two feature spaces A and B with dimensions of $T_1 \times D \times k \times k$. The original input data is encoded and pooled using a transformation technique.

Spatial Attention: In the spatial domain, each pixel in the feature map correlates with all other pixels. The spatial feature map with dimensions $T_1 \times k \times k \times D$ is transformed into feature tensors $T_1 \times D \times k \times k$ by applying a reshaping operation. The spatial domain relies on attention weights, α, calculated by the inner product of two encoded feature transformation vectors, $A(I)$ and $B(I)$ as Equation (1), where the $\mathrm{Softmax}$ normalization function is used on each channel to establish relationships between pairs of positions as shown in Equation (2).

$$r_{ij} = A(I_i)^T B(I_j) \tag{1}$$

$$\alpha_{i,j} = \mathrm{Softmax}(r_{ij}) \tag{2}$$

Where $(k \times k)$ the total amount of pixels in a spatial feature map. The results is spatial correlation matrix with dimension $\alpha \in \mathbb{R}^{(k \times k)}$.

Depth Attention: This module is utilized to compute temporary connections between pixels. We construct an explicit depth attention module to capture temporal features related to other feature maps. The structure of the depth attention module is illustrated in Figure 2. In contrast to spatial attention, depth attention computes a similarity matrix using a temporal approach. Specifically, we transform the spatial feature map with dimensions $T_1 \times k \times k \times D$ into normalized $((T_1 \times D) \times k \times k$ after softmax, where the output is a temporal correlation matrix of size $\beta \in \mathbb{R}^{(T_1 \times D)}$. The aggregated feature maps operate stack-wise to compute the feedback of both spatial and depth attention, taking $[\alpha, \beta]$ as input. The 3D

convolution follows channel-wise correlation to allow the network to reduce the depth of the concatenated spatial-depth correlation data. In this study, the number of 3D filters, D', equals D. Finally, the output of I' is element-wise multiplied with the input feature map, I, to generate the output of $\bar{I}$.

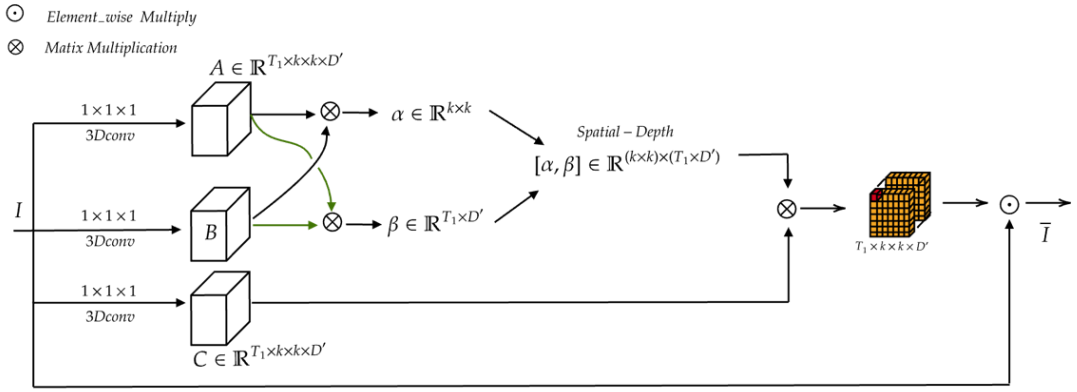

Figure 2. The proposed SBN

3.2. Proposed Model DNN-2

In Fig 1(b), To improve recognition performance, we provide the overall architecture together with the particular attention mechanisms based on the fusion of multiple effective action clues across the spatial and temporal domains. The input data are the resampled RGB and optical-flow channel streams where the motion maps contribute to the attention. First, two streams feed into the HeadSBN layer to produce the spatial-depth attention vector. Second, the Deep CapsNet classifier employs multiple CapsNets layers to produce many capsules from these feature vectors and then fulfills the classification.

3.2.1. Spatial-Motion feature exaction by 3DCNN

For training, the transfer learning technique is employed to ensure the most effective training of the model. This is achieved by leveraging knowledge transferred from related tasks. Such a strategy has the potential to provide our architecture with highly effective initial configurations for the learning process on the network's multi-modal cost function. This is because the datasets used in the proposed DNN may not be rich or diverse. This transfer learning strategy can reduce the amount of time required to train the proposed network by using fewer training samples to achieve the desired level of performance. Therefore, we utilize the pre-trained Inception I3D model, fine-tuned on ImageNet and Kinetics, as the foundation for our CNN. Subsequently, these CNNs are retrained using our dataset to build spatial and motion feature maps. In Figure 1, the I3D module up to 'mixed-4f' generates feature maps extracted from synthesized layers at the size of $T_1 \times k \times k \times D$, where T_1, D, and $k \times k$, respectively, represent temporal resolution, object channel size, and spatial resolution of an object map. This fundamental feature is the dense spatial-temporal representation of the video stream being input.

3.2.2. Top-Heavy CapsNet (THC)

Owing to research the effect of CapsNet in our DNNs architecture. As shown in Fig. 1(b), we designed the structure of Top-Heavy CapsNet (THC). A major advantage gained by DCapsNet is using a concatenation of different scales on different temporal dimensional vectors for the primary capsule. The semantic information in both M^s, M^f are obtained by the feature extraction from the 'mixed_4f' convolution layer of the I3D, which generated in D feature map of dimension $T_1 \times k \times k$. The feature maps are employed to extract the difference scale feature vector dimension of capsules composed of 32 capsule types by convolution operation, with ReLU and sigmoid activations.

For the UCF101 dataset example, the $T_1 = 30$, M_p^s with k=14 which result in D=1664 feature map of dimension $30 \times 14 \times 14 \times 1664$. Since the multi-dimensional primary capsule is obtained, 1,152 ($6 \times 6 \times 32$) primary capsules with 8D vector. The features are used to create a convolution layer. Kernel: $7 \times 7 \times 7$, stride $=(2 \times 1 \times 1)$ is to be output

shape $12 \times 8 \times 8 \times 32$. This is followed by a 3D convolution capsule layer with 32 capsules types with the kernel: $5 \times 3 \times 3$, stride $=(1 \times 1 \times 1)$ to be the output of the dimension of $8\times 6 \times 6 \times 32$. The last 3D convolution capsule operation with kernel $1\times 1 \times 1$ and stride $=(2 \times 1 \times 1)$ to get the different scale of the primary capsule of $4 \times 6 \times 6 \times 32$. In the third stage, dedicated to the final classification layer (class capsules), a convolutional capsule layer is fully connected to L capsules in which L represents the number of action classes. The capsule output is determined by Caps-Pool, with coupling coefficients adapted through an iterative EM routing process. This process dynamically updates the probabilities of the class output neurons, akin to the following [16].

4. Results and Discussion

4.1. Dataset and Configuration for Experiments

To compare the proposed and traditional DNNs, the three datasets UCF101, HMDB51, and Traffic Police (TP) are used. The UCF101 dataset comprises 13,320 video clips, each corresponding to one of 101 distinct actions. These actions encompass five categories: human-object interaction, body movements, human interaction, playing musical instruments, and sports. The HMDB51 dataset contains video clips that correspond to 51 different activity categories. These video clips have a wide variety of backgrounds and contexts, as well as different camera movements. The data partitioning configurations supported by the original UCF101 dataset and the HMDB51 dataset were used for training/validating and testing. The Traffic Police (TP) dataset [1] is utilized, consisting of 21 video clips. These clips have a frame resolution of 1080×1080 pixels and a frame rate of 15Hz.

Similar to [3], at a frame rate of 30 frames per second (fps) for UCF101, 25 fps for HMDB51, and 30 fps for TP, all video clips were divided into overlapping segments, each lasting 3 seconds, indicating T=90 and 75, respectively. There is a 50% overlap between consecutive video segments. In these datasets, each video clip represents only one classified action. Once the video segments are prepared, they are preprocessed to become input segments with the corresponding number of frames, T_1=30 (HMDB51 and TP) and T_1=25 (UCF101). Additionally, the frames of these input segments are resized to 224 x 224 to match the input format specifications of the proposed DNN. With τ=3, we generate T_1-sized 224×224×3 RGB and 224×224×2 optical flow channels from the T RGB images of the video segment, similar to the procedure in [3].

The goal is to recognize actions in videos. The TP dataset has 21 video clips associated with nine actions, each with a frame resolution of 1080×1080 pixels and a frame rate of 15Hz. The video clips are split into training, validation, and testing sets. The model is trained for 75 epochs, and training stops if accuracy does not consistently increase over ten consecutive epochs. Training stopped by early stopping after ten epochs is approached. The training, validation loss, and accuracy are recorded for each epoch, and the model with the best performance is selected for testing. Optimization is performed using the Adam optimizer with a momentum of 0.9, and the learning rate is 1e-5, decay=1e-6, and the random gradient descent is batch-wise reduced. T_1 represents the number of frames accessed from a video clip, and our computational platform consists of an 8-core Intel Core i7 CPU and two Nvidia Titan X 1080/12 GB RAM GPUs @ 32GB. The CPU system runs on Ubuntu 18.04 64-bit, and the Anaconda Python distribution provides GPU support. Our model is fully trainable end-to-end, allowing it to be used with larger and more complex datasets. We developed our models in TensorFlow and provided both the code and our pre-trained models publicly. We also utilize model checkpointing to save the best-performing model, the model with the lowest validation loss, and then use that model for final predictions. The frames in the video segments are resized to 224x224px. We consider analyzing the temporal length changes of video segments (T_1=10) at one-time step to be sufficient for accurate activity predictions. For this, we use a pre-trained model to make initial predictions. We then concatenate these feature maps into a single sample, which is the input to our neural regression network. Finally, we obtain the system's final classification based on the inference layer.

Evaluation metrics play an important role in estimating how well the optimal classifier occurred during the building model in terms of action recognition. In the study, the metrics of average accuracy and confusion matrix approach for an experiment on one vs all classification techniques for multi-classification as similar[2], [3]. The sum of operations in all

convolution layers is then computed based on the number of output feature maps, kernel sizes, input channels, and output channels.

4.2. Effect of Spatial-Depth Attention in the Proposed Model DNN-1

To illustrate the impact of the SBN layer, we designate the proposed DNNs without and with the SBN layer as Type-1, Type-2, and Type-3 DNNs. Type-1 combines 3DCNN+RL, Type-2 combines 3DCNN+SA+RL, and Type-3 combines 3DCNN + SBN. Table 1 summarizes the accuracy of eight actions detected by these two types of DNNs. Test accuracy for the "Stop" and "Moving Straight" categories is approximately 99.9 % when using Type-3 DNN, while the accuracy for the "Right Turn" class is around 90% when using Type-1 DNN due to numerous misclassifications. These findings indicate that the attention mechanism included in the SBN layer enables exceptional performance across all actions. We present the outcomes of multi-class classification on the traffic police dataset. Notably, certain categories such as "stops" and "moving" actions exhibit accuracy levels close to 100%, whereas the accuracy for identifying classes like "right-turn" hovers around 93.6%, owing to some misclassifications. The best classifier in the proposed DNN is showcased, highlighting instances where the model faltered in recognizing most of the False Negatives in the "Right Turn" class. Conversely, activities such as Moving Straight, Stop, Slow Down, and Pull Over were correctly predicted, with the majority achieving accuracy levels surpassing 93% in our proposed architecture.

Table 1. Accuracies of 8 action recognition by Type-1, Type-2, and Type-3 DNN model

Categories	Average accuracies (%)		
	Type-1((re-trained from [2]))	Type-2 (re-trained from [2])	Type-3
Stop	97.5	99.1	99.9
Moving Straight	95.9	99.6	99.9
Left Turn	92.1	92.5	95.4
Left Turn Waiting	95.5	95.6	96.1
Right Turn	90.4	93.5	93.7
Lane Changing	94.2	97.9	97.0
Slow Down	97.1	99.8	99.3
Pull Over	95.0	99.1	99.5

Confusion matrices serve as an effective means of data visualization, adeptly presenting comprehensive results that encapsulate correct classification accuracy and misclassification details for each predicted category. Our experimental findings offer an in-depth analysis of our framework's performance in the context of action recognition tasks, aligning with the ground truth.

Table 2 presents performance metrics on the TP dataset, with the accuracy of Type-2 DNN increasing by 2.4% compared to Type-1 DNN. Notably, Type-3 exhibits a significant improvement compared to Type-1 and Type-2, with increases of 2.9% and 0.5%, respectively. This improvement is attributed to Type-2 utilizing spatial attention and Type-3 employing a dual attention mechanism.

Table 2. Performance metrics of three Types on the TP Dataset

DNNs	Training performance				Average F1-Measure		Test acc
	Training acc	Validate acc	Training Loss	Validate loss	Micro	Macro	
Type-1 (re-trained from [2])	99.8%	94.5%	0.09	0.10	0.93	0.94	94.7%
Type-2 ((re-trained from [2]))	99.9%	98.9%	0.01	0.05	0.97	0.96	97.1%
Type-3	99.9%	99.3%	0.009	0.02	0.97	0.98	97.6%

4.3. Analyses and Comparisons of Experimental Results in the proposed DNN-2

At first, partial structures from the proposed DNN were evaluated to understand their performance contributions. After that, the optimized DNN setup was confirmed. The experimental outcomes of the proposed and conventional DNNs are compared and analyzed.

4.3.1. Exploration of attention mechanisms versus performance on comparison of Individual CapsNet structure

During the run of the experiments, the various types of DNNs used the RGB stream with or without the motion stream, yielding the results in Table 3. Compared to the 3DCNN with SBN and the 3DCNN with only RGB input, the attention mechanism indeed increases the accuracies by 1.6% and 1.1% at the UCF101 and HMDB51 datasets, respectively. The outcomes from the I3D+SBN+ DRCapsNet and the type without CapsNet (I3D+SBN + FC) reveal that the deep DRCapsNet contributes to the accuracies lifted. Using the AMS to gate OF and RGB, the I3D+AMS+ DCapsNet outperforms the I3D + DCapsNet by 7.5% and 8.9% on the UCF101 and HMDB51, respectively. Moreover, the performance of I3D+AMS+DCapsNet in the downstream task surpasses that of other DNNs, exhibiting an average accuracy increase ranging from 2.3% to 16.9%. This notable improvement can be attributed to the efficacy of attention mechanisms, specifically the Top-heavy CapsNet, in providing meaningful static and dynamic information related to the subject's body joints during an activity. The integration of multiple streams contributes to enhanced performance by leveraging complementary feature information derived from the subject's appearance and motion. This information can be obtained when these multiple streams are available.

Table 3. The performance-based-I3D transfer learning with different CapsNet in sigle and due stream input of UFC101 and HMDB51 dataset.

Model	Classification Accuracies (%)		
	Input	UCF101	HMDB51
I3D + MLP	RGB	95.32	74.44
I3D + NA + MLP	RGB	96.61	75.86
I3D + SA + DCapsNet	RGB	96.85	75.94
I3D + MLP (class score fusion)	RGB+ OF	97.88	79.77
I3D + NA + MLP (class score fusion)	RGB+ OF	98.02	79.99
2I3D + SBN + DCapsNet	RGB+ OF	98.17	80.43

4.3.2. DNN_2 addressed by single or two input streams

The proposed DNN were designed as an integrated version of two DNNs that look after the RGB streams as well as the optical-flow channel streams. In light of this, three other topologies that were produced from the proposed DNN by utilizing the CapsNet classifiers are being looked into. These three use the apparent streams, the optical-flow channel streams, and the proposed DNN using two streams. The experimental results associated with classification accuracies are summarized in Table 4. The accuracy of the suggested DNN, which manages two streams instead of just one, is much better than that of DNNs that only analyze one input stream, with an increase ranging from 0.6% to 4.6%. In most instances, the higher structure of our DNN produces superior results to those of the lower structure. However, as a result, the specific information associated with the subject motion plays a crucial role in determining whether or not an action was taken. As a consequence, the motion model displays a better degree of accuracy than the Appearance model.

Table 4. Classification accuracies of the appearance, motion, and complete structure of our model DNN.

Model	Classification Accuracies (%)			
	Input	UCF101	HMDB51	TP
Appearance	RGB	95.32	74.44	97.7
Motion	OF	97.01	76.84	98.0
Appearance & Motion	RGB+ OF	98.04	80.32	98.1

4.4. Discussion and Visualization

Figure 3 depicts the action recognition process for three activities, showcasing the attention feature maps and corresponding emphasized action areas. Warm colors indicate a high score, directing attention to crucial aspects of action recognition. The proposed DNN effectively highlights significant motion patterns of various subjects and integrates them into a discriminative representation for accurate action recognition. The operational flow depicted in Fig. 1 is dissected so that the outputs from the intermediate levels can be visualized, and then the contributions of those layers can be found. This is done so that the proposed DNN can be comprehended. The effectiveness of spatial attention is illustrated by the figures in Figure 3, which highlight the intensity of the region. In each illustration, some regions of the picture are denoted by using warm colors to represent the attention locations and the intensities. When the intensity is higher, the color seems to be more vibrant.

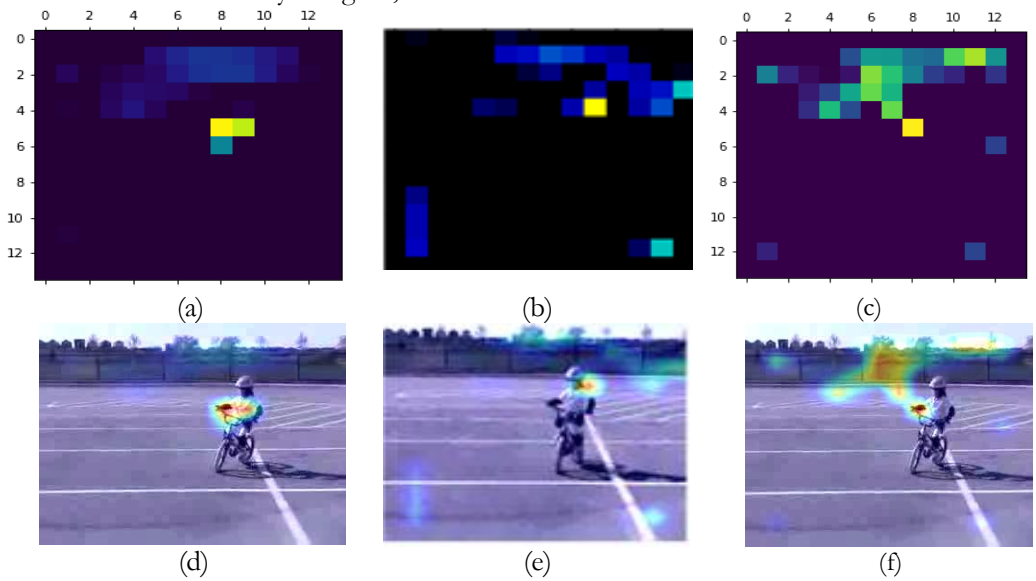

Figure 3. Example of visualization of the output attendant feature from the intermediate layers on UFC101: (**a**) the feature map from I3D; (**b**) feature map with the aid of NA; (**c**) the feature Map with the aid of SNB; (**d**)mask of the region with high intensity from I3D, (**e**) mask of the region with high intensity from NA; (**f**) mask of the region with high intensity from SNB. Warm is a strongly involved region of action behavior.

Figures 3(a) and (d) visually represent the feature map obtained by 3DCNN. The highlighted area is the tiny region obtained by focusing on the topic. The output of the proposed NA layer is shown in Figures 3(b) and (e), which simply require the spatial feature map as an input. Figures 3(c) and (f) reveal that with the use of the motion feature maps, attention in the SBN layer is brought to more significantly concentrated regions associated with subject bodies, ultimately leading to correct action prediction. Notably, in the context of attention heat maps that are indicated by a variety of colors, the method that we have presented, which takes subject motion into account, does actually improve recognition based on meaningful and distinguishing characteristics associated with subject activities. Based on the attention-based spatial-temporal relationships from the appearances and motions of subjects, the attention mechanisms adopted in the proposed DNN exhibit outstanding performance, as can be seen from the visualization, the steps broken down, and the comparisons made.

5. Comparison

5.1. Comparison state of the art on police traffic

Table 5 presents a comparative analysis of the performance between the proposed and previous DNN models using the identical traffic police video dataset. The experimental findings reveal that the proposed DNN, excluding the recurrent network, achieves a superior accuracy improvement of 4.3% compared to conventional DNNs. Moreover, the proposed DNN demonstrates a marginal accuracy enhancement of 0.1%. Notably, this study exclusively utilizes RGB streams without incorporating skeletal postures, optical flow, or

comparable techniques. Consequently, the proposed DNN emerges as a potent tool for facilitating action recognition across diverse content-aware applications.

Table 5. Average accuracies of the proposed model and conventional DNNs

DNNs	Years	Features	Accuracies (%)
[23]	2020	Pose + handcrafted features	93.3
[24]	2020	Pose graph	97.5
[3]	2021	Pretrained ImNet+Kinetics R_FSRH_D	98.1
Proposed DNN		RGB, Pretrained ImNet+Kinetics HeadSBN+RL	97.6
		RGB+OF, Pretrained ImNet+Kinetics HeadSBN+DCapsNets	98.1

5.2. DNN_2 Comparisons with the SOTA in RGB input

Table 6 outlines the mean accuracies achieved by both the proposed and traditional DNNs when utilizing RGB input datasets from UCF101 and HMDB51. First, in comparison with I3D and S3D, our model plus SBN and training combined with deep CapsNet are superior to the margin by 1.3% and 0.1%, respectively.

Table 6. The average accuracies of comparison on the UCF-101 and HMDB-51 dataset with RGB input

DNNs	Years	Pre-train	UCF101	HMDB51
Res3D [25]	2017	Sports-1M	85.5	54.9
TSN [26]	2016	Kinetics	85.7	-
I3D [5]	2017	ImNet+Kinetics	95.6	74.8
S3D [27]	2018	ImNet+Kinetics	96.8	75.9
[3]	2021	ImNet+Kinetics	96.9	76.1
I3D+SBN+DCapsNet		ImNet+Kinetics	97.1	76.1

The multiple stream conventional model integrated with I3D manner, Improved Dense Trajectory (IDT), plus pose motion, RGB, and Optical flow together yielded relatively good performance [3], [8], [12], [13]. However, effectively learning spatial appearances and temporal motion behaviors within the complex contexts of videos remains a challenging task. According to Table 7, the performance of the proposed DNN is comparable to that of the best one at UCF101. The reasons for this are that the relevant regions, both temporal and

Table 7. Performance comparison of the proposed and conventional model DNNs using HMDB51 and UCF101 datasets

DNNs	Years	Characteristics	UCF101	HMDB51
[21]	2018	I3D + Pose motion, pre-training	98.2	80.9
[5]	2019	Motion-augmented, pre-training	98.1	80.9
[20]	2019	LGD-3D RGB+OF	98.2	80.5
[8]	2019	Hallucinating IDT and I3D OF	-	82.5
[11]	2020	BubbleNET	97.6	82.6
[13]	2020	SlowOnly-8x8-R101 + Flow	98.6	83.8
[9]	2021	VidTr	96.7	74.4
[3]	2021	RGB+OF+Tske, I3D ImageNet+Kinetics pre-trained, CapsNet	98.5	82.1
[6]	2022	BQN (TSM R50)	97.6	77.6
[14]	2022	Temporal Squeeze Network	95.2	71.5
[12]	2022	PERF-Net (Kinetics-600 pretrain)	98.2	82.0
Our DNN		ImNet+Kinetics pre-trained of top-heavy DeepCapsNet	**98.6**	80.4

spatial, are emphasized adequately with the assistance of temporal features, motion features, and spatial features rather than IDT and I3D. Specifically, the multiple CapsNets within the proposed classifier translate these feature vectors into extensive capsules to achieve refined classifications and enhance overall performance.

6. Conclusions

In this paper, we address the challenge of constructing a highly efficient CapsNet by combining it with an attention mechanism and integrating it into a 3DCNN for video recognition tasks. We introduce a novel Top-Heavy CapsNet architecture incorporating special attention mechanisms, leveraging fused and effective action cues across temporal and spatial domains to enhance identification performance. The input data are the resampled RGB and optical-flow channel streams where the OF maps contribute to the attention. First, these two streams feed into the corresponding HeadSBN-based 3DCNN backbone to yield the spatial and motion feature maps. Second, the deep CapsNet employs multiple 3D convolutions Capsule layers to produce capsules class, then fulfills the classification. We have conducted an empirical investigation into the impact of various spatiotemporal convolutions on video action recognition. Our model outperforms the same performance on UCF101 in the RGB domain.

Funding: This research received no external funding

Acknowledgments: This research is supported by the International School, Vietnam National University, Hanoi (VNU-IS).

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] M.-H. Ha and O. T.-C. Chen, "Action Recognition Improved by Correlations and Attention of Subjects and Scene," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, Dec. 2021, pp. 1–5. doi: 10.1109/VCIP53242.2021.9675340.

[2] M.-H. Ha and O. T.-C. Chen, "Deep Neural Networks Using Residual Fast-Slow Refined Highway and Global Atomic Spatial Attention for Action Recognition and Detection," *IEEE Access*, vol. 9, pp. 164887–164902, 2021, doi: 10.1109/ACCESS.2021.3134694.

[3] M.-H. Ha and O. T.-C. Chen, "Deep Neural Networks Using Capsule Networks and Skeleton-Based Attentions for Action Recognition," *IEEE Access*, vol. 9, pp. 6164–6178, 2021, doi: 10.1109/ACCESS.2020.3048741.

[4] K. Rajesh, V. Ramaswamy, and K. Kannan, "Prediction of Cyclone Using Kalman Spatio Temporal and Two-Dimensional Deep Learning Model," *Malaysian J. Comput. Sci.*, pp. 24–38, Nov. 2020, doi: 10.22452/mjcs.sp2020no1.3.

[5] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-Augmented RGB Stream for Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7874–7883. doi: 10.1109/CVPR.2019.00807.

[6] G. Huang and A. G. Bors, "Busy-Quiet Video Disentangling for Video Classification," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 756–765. doi: 10.1109/WACV51458.2022.00083.

[7] G. Huang and A. G. Bors, "Learning Spatio-Temporal Representations With Temporal Squeeze Pooling," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 2103–2107. doi: 10.1109/ICASSP40776.2020.9054200.

[8] L. Wang, P. Koniusz, and D. Huynh, "Hallucinating IDT Descriptors and I3D Optical Flow Features for Action Recognition With CNNs," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 8697–8707. doi: 10.1109/ICCV.2019.00879.

[9] Y. Zhang *et al.*, "VidTr: Video Transformer Without Convolutions," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 13557–13567. doi: 10.1109/ICCV48922.2021.01332.

[10] R. K. Rachman, D. R. I. M. Setiadi, A. Susanto, K. Nugroho, and H. M. M. Islam, "Enhanced Vision Transformer and Transfer Learning Approach to Improve Rice Disease Recognition," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 446–460, Apr. 2024, doi: 10.62411/jcta.10459.

[11] B. Igor L. O., M. Victor H. C., and W. R. Schwartz, "Bubblenet: A Disperse Recurrent Structure To Recognize Activities," in *2020 IEEE International Conference on Image Processing (ICIP)*, Oct. 2020, pp. 2216–2220. doi: 10.1109/ICIP40778.2020.9190769.

[12] Y. Li, Z. Lu, X. Xiong, and J. Huang, "PERF-Net: Pose Empowered RGB-Flow Net," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 798–807. doi: 10.1109/WACV51458.2022.00087.

[13] H. Duan, Y. Zhao, Y. Xiong, W. Liu, and D. Lin, "Omni-Sourced Webly-Supervised Learning for Video Recognition," 2020, pp. 670–688. doi: 10.1007/978-3-030-58555-6_40.

[14] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, S. Xia, and Y.-J. Liu, "Motif-GCNs With Local and Non-Local Temporal Blocks for Skeleton-Based Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2009–2023, Feb. 2023, doi: 10.1109/TPAMI.2022.3170511.

[15] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-Attention Generative Adversarial Networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, vol. 97, pp. 7354–7363. [Online]. Available: https://proceedings.mlr.press/v97/zhang19d.html

[16] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=HJWLfGWRb

[17] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne, and R. Rodrigo, "DeepCaps: Going Deeper With Capsule Networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 10717–10725. doi: 10.1109/CVPR.2019.01098.

[18] W. Wang, F. Lee, S. Yang, and Q. Chen, "An Improved Capsule Network Based on Capsule Filter Routing," *IEEE Access*, vol. 9, pp. 109374–109383, 2021, doi: 10.1109/ACCESS.2021.3102489.

[19] W. Huang and F. Zhou, "DA-CapsNet: dual attention mechanism capsule network," *Sci. Rep.*, vol. 10, no. 1, p. 11383, Jul. 2020, doi: 10.1038/s41598-020-68453-w.

[20] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified Spatio-Temporal Attention Networks for Action Recognition in Videos," *IEEE Trans. Multimed.*, vol. 21, no. 2, pp. 416–428, Feb. 2019, doi: 10.1109/TMM.2018.2862341.

[21] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion Representation for Action Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7024–7033. doi: 10.1109/CVPR.2018.00734.

[22] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, vol. 37, pp. 2048–2057. [Online]. Available: https://proceedings.mlr.press/v37/xuc15.html

[23] J. He, C. Zhang, X. He, and R. Dong, "Visual Recognition of traffic police gestures with convolutional pose machine and handcrafted features," *Neurocomputing*, vol. 390, pp. 248–259, May 2020, doi: 10.1016/j.neucom.2019.07.103.

[24] Z. Fang, W. Zhang, Z. Guo, R. Zhi, B. Wang, and F. Flohr, "Traffic Police Gesture Recognition by Pose Graph Convolutional Networks," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, Oct. 2020, pp. 1833–1838. doi: 10.1109/IV47402.2020.9304675.

[25] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "ConvNet Architecture Search for Spatiotemporal Feature Learning," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1708.05038

[26] L. Wang *et al.*, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *Computer Vision – ECCV 2016*, 2016, pp. 20–36. doi: 10.1007/978-3-319-46484-8_2.

[27] D. Zhang, X. Dai, X. Wang, and Y.-F. Wang, "S3D: Single Shot multi-Span Detector via Fully 3D Convolutional Networks." Jul. 20, 2018. [Online]. Available: http://arxiv.org/abs/1807.08069