Sparse Dense Transformer Network for Video Action Recognition

Xiaochun Qu, Zheyuan Zhang, Wei Xiao, Jinye Ran, Guodong Wang,
and Zili Zhang(✉)

College of Computer and Information Science,
Southwest University, Chongqing 400715, China
zhangzl@swu.edu.cn

Abstract. The action recognition backbone has continued to advance. The two-stream method based on Convolutional Neural Networks (CNNs) usually pays more attention to the video's local features and ignores global information because of the limitation of Convolution kernels. Transformer based on attention mechanism is adopted to capture global information, which is inferior to CNNs in extracting local features. More features can improve video representations. Therefore, a novel two-stream Transformer model is proposed, Sparse Dense Transformer Network(SDTN), which involves (i) a Sparse pathway, operating at low frame rate, to capture spatial semantics and local features; and (ii) a Dense pathway, running at high frame rate, to abstract motion information. A new patch-based cropping approach is presented to make the model focus on the patches in the center of the frame. Furthermore, frame alignment, a method that compares the input frames of the two pathways, reduces the computational cost. Experiments show that SDTN extracts deeper spatiotemporal features through input policy of various temporal resolutions, and reaches 82.4% accuracy on Kinetics-400, outperforming the previous method by more than 1.9% accuracy.

Keywords: Transformer · Action recognition · Two-stream · Frame alignment · Patch crop

1 Introduction

The development of computer technology has been applied in all aspects of life [4–6,16,21,24]. With the diversification of content presentation forms on social platforms, videos have progressively risen to prominence in our lives. Millions of videos are published on YouTube, TikTok, and other platforms on a daily basis. Thus, understanding and analyzing the content of videos play a critical role in video sharing and monitoring fields. Similarly, the explosive growth of video streams has also posed a challenge to today's video field research: how to achieve high-precision video understanding under limited computational cost?

To reduce the computational task, a common practice in action recognition is generally used to sample specific frames from video, feed the sampled

G. Memmi et al. (Eds.): KSEM 2022, LNAI 13369, pp. 43–56, 2022.
https://doi.org/10.1007/978-3-031-10986-7_4

frame into the designed network, and finally perform action recognition [12]. The original intention of using a specific method to sample frames is to reduce computational cost and redundant frames, making the network more suitable for long-term motion. However, the details of changes between consecutive frames will be inevitably ignored, as with random sampling and sparse sampling. The latest two-stream methods [11,17] used different time resolutions for two pathways, which avoids the problem of missing essential frames and extracts features more efficiently. Therefore, the two-stream method is more extensively utilized in action recognition Convolutional Neural Networks.

The traditional two-stream method [27] feeds optical flow and RGB separately for action recognition. Meanwhile, TSN [33], one of the variants of two-stream, provided a novel idea to process long-term action recognition tasks by segmenting in the temporal dimension. Of course, research on action recognition is not limited to 2D, and even excellent results have been obtained using 3D CNNs [18,20]. 3D CNNs will undoubtedly extract more spatiotemporal features than 2D CNNs [28]. There are two types of cells in the primate visual system, Parvocellular (P-cells) and Magnocellular (M-cells). Among them, M-cells are interested in rapid time changes, and P-cells are sensitive to significant spatial features [31]. SlowFast [11] obtained accurate spatiotemporal information more efficiently by imitating two cells. However, due to the limitation of the size of the convolution kernel, the two-stream network based on CNNs often cannot effectively model the global features, resulting in the lack of feature diversity.

Transformer achieved excellent results in the field of Natural Language Processing (NLP) [32]. In order to apply the Transformer to images, ViT [8] regarded each image as consisting of many 16×16 patches. Video and NLP, as compared to image, have a higher level of similarity, whether sequential or logical [26]. Therefore, the Transformer used in image is also suitable for video and can even achieve better performance in video research. However, when Transformer processes images, it often ignores the intrinsic structure information inside each patch, resulting in the lack of local features [14].

To shed new light on studying the applicability of CNNs architecture on Transformer, a two-stream Transformer architecture is proposed to combine local and global features more effectively. Simultaneously, to achieve precision and speed trade-off, we execute frame alignment operation to ensure that the Sparse pathway's input frames are the same as the Dense pathway's input frames. Thus, the frames feeding the Dense pathway do not need to be processed all at once, reducing the Dense pathway's computation cost. Patch crop, a new cropping method, is also designed to focus on the center of videos.

Our contributions can be summarized as follows:

- A novel architecture, **S**parse **D**ense **T**ransformer **N**etwork(SDTN), is proposed to combine two-stream method and Transformer, considerably enhancing action recognition performance.
- A new cropping approach with 16×16 as the basic unit, Patch crop, which allows the network to pay more attention to the patches in the center of videos.

- Experiments demonstrate that SDTN achieves accuracy improvements on Kinetics-400 and presented Light kinetics-400. Additionally, the frame alignment is leveraged as information aggregation method to improve the performance of SDTN.

In the remainder of the paper, we first introduce related work in Sect. 2. In Sect. 3, we illustrate the proposed model. Section 4 and Sect. 5 show the experimental results and analyze based on them. Finally, the conclusion and outlook are presented in Sect. 6.

2 Related Work

2.1 CNNs in Action Recognition

CNNs have long been the standard for the backbone architectures in action recognition. Simonyan and Zisserman proposed two-stream architecture, whose method is to combine RGB and optical flow for action recognition [27]. To some extent, the computational cost of the optical flow is relatively expensive, and researchers continue to do further research on this basis. Subsequently, TSN [33] offered an exciting solution to advance our knowledge of long-term video action recognition tasks: it segments the video into N clips in the time dimension; then, each clip is input into the two-stream network, which is effectively improved the accuracy of the action recognition in long-term videos. ECO [35] provided the ECO-Full network based on TSN, which is also a parallel of two networks.

For 3D CNNs, C3D [28] is a pioneering work that has designed an 11-layer deep network. R3D can be regarded as a combination of two outstanding research of Resnet and C3D [15,30]. I3D [7] extended 2D CNNs to 3D CNNs and showed that 3D CNNs extract features substantially more effectively than 2D CNNs. Because 3D CNNs have a lot of parameters and high computational costs, researchers are trying to lower their computational complexity to that of 2D CNNs. As a result, P3D [25] and R(2+1)D [30] have been trying to replace 3D CNNs with 2D CNNs and have achieved good results on large-scale datasets. V4D [34] achieved accuracy improvements by adding the clip dimension to 3D CNNs. The SlowFast networks [11], inspired by the biological research of retinal ganglion cells in the primate visual system, found that combining multiple time resolutions is helpful for accurate and effective action recognition in experiments.

2.2 Transformer in Action Recognition

Transformer initially achieved excellent performance in NLP, and later it was introduced into computer vision.

Rohit Girdhar et al. [13] used Transformer structure to add the information before and after the video to the final vector for classification and positioning. Then, Vision Transformer (ViT) [8] proposed a brand new idea: can continuous patches represent an image? In other words, ViT decomposes an image into 16×16 patches. Since then, ViT realized the transformation of the backbone of computer vision from CNNs to Transformer. The great success of the image Transformer has also led to the research on the architecture of the action recognition task based on the Transformer. VTN [23] proposed to add a temporal attention encoder to pre-trained ViT, which performed well on action recognition datasets. ViViT [1] was based primarily on ViT and tried to solve the video task completely using the Transformer architecture. ViViT studied four spatial and temporal attention factor designs of the pre-trained ViT model, and recommended a VTN-like architecture. The latest Video Swin Transformer [22] has made the number of tokens less and less through multiple stages, and the receptive field of each token has increased. So the computational cost of the Video Swin Transformer is reduced, while the precision and speed are both improved.

3 Sparse Dense Transformer Network

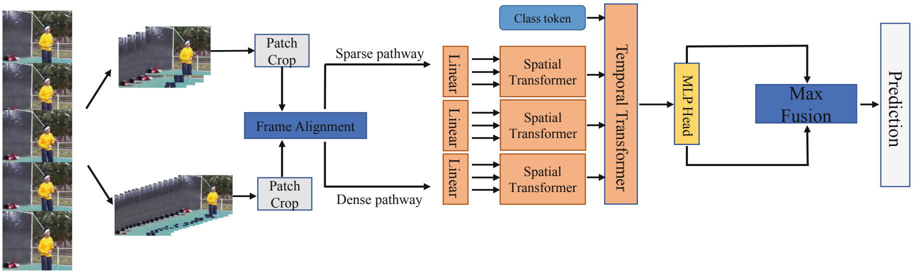

Fig. 1. Sparse Dense Transformer Network processes information with two different time resolutions. The Sparse pathway and the Dense pathway carry out patch crop operation simultaneously. After that, the frames of the two pathways undergo frame alignment and then enter their respective networks. Finally, through max fusion, the video-level prediction results are obtained.

Following the STAM [26], we use the idea of SDTN to simulate the P-cells and M-cells of the biological vision system. As shown in Fig. 1, SDTN can be described as a single stream structure with two different time resolutions as input. The Sparse pathway samples informative frames with lower temporal resolution. The Dense pathway is sampled at high temporal resolution and is sensitive to rapid temporal changes. Meanwhile, we explore how SDTN can benefit from various time resolutions, motivated by Coarse-Fine networks. Moreover, based on the intuition that the action is more concentrated in the center of the video, SDTN adopts a patch-based cropping method to make the pathway focus more on the central patches of the input frame.

In SDTN, we are confronted with two major challenges: (i) how to improve the network's accuracy while reducing the Dense pathway's computational cost; (ii) how to effectively integrate the information of the Sparse pathway and the Dense pathway. First of all, to minimize the Dense pathway's computational cost, we propose frame alignment, which is a method of ensuring that the frames sampled by the Sparse and Dense pathways are consistent, therefore reducing the amount of computing on the Dense pathway. Furthermore, to assure the accuracy of SDTN, we design different degrees of information fusion experiments to contrast the impact of various methods of fusion.

3.1 Frame Alignment

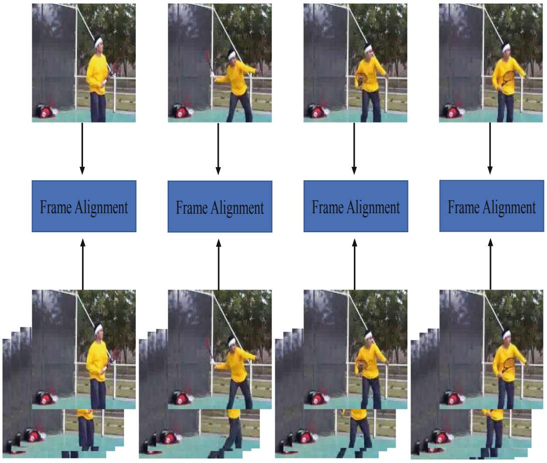

Fig. 2. Frame Alignment compares the sampled frames of the Sparse pathway and the Dense pathway. In this way, the consistency of the processing information of the Sparse pathway and the Dense pathway can be guaranteed. While ensuring the diversity of network input information, it also reduces the computational cost of the Dense pathway.

Specifically, SDTN aims to design a new two-stream Transformer architecture while maintaining the original network features to show a speed-accuracy trade-off. The Sparse and Dense pathways sample various amounts of frames from the whole video. The input to each pathway is $X \in \mathbb{R}^{H \times W \times 3 \times F}$ consists of F RGB frames of size $H \times W$ sampled from the video. During the learning process, each pathway derives the preliminary inference results and finally applies fusion approach to produce the video-level inference.

We expect the Dense pathway to assure excellent accuracy while operating with a minimum computational cost. Therefore, we design frame alignment on the original foundation to ensure the variety of the two networks' input while maintaining the consistency of the features.

As shown in Fig. 2, the Sparse pathway of SDTN samples 16 frames, and the Dense pathway samples 64 frames. Before input, we align the input frames of two networks according to the frame id. The frames of the Dense pathway are compared with the frames of the Sparse pathway before the input to determine whether the two networks' input is consistent. Accordingly, we reduce the input into the Dense pathway following this comparison, reducing the computational cost of SDTN.

3.2 Patch Crop

There should be a subject, a predicate, and an object in a sentence. If the sentence is complicated, then there are attributives and complements. When the Transformer processes a frame, it treats a frame as a complete sentence, with each patch representing a word. To this end, we try to adapt essential human thinking to the computer in this experiment. For example, when processing a sentence, we give greater attention to the subject, predicate, and object. Similarly, we believe that subjects, predicates, and objects in the image are made up of patches. Then, based on the intuition that the behavior mainly occurs in the center of videos, we consider the patches in the center of videos as the subject, predicate, and object of this frame.

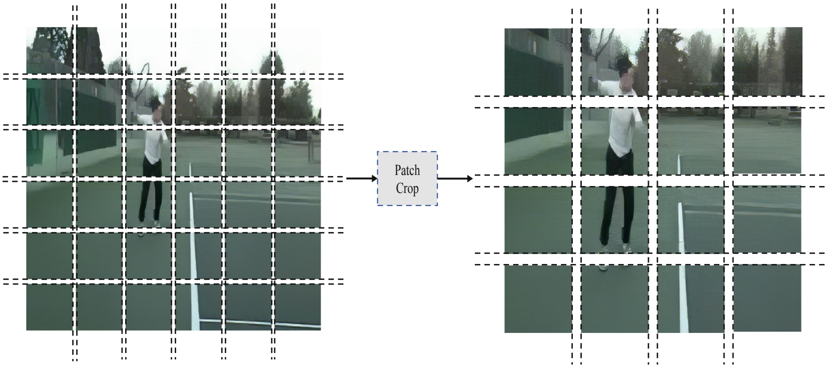

Fig. 3. Patch Crop. Sparse Dense Transformer Network is expected to focus more on the frame's center. To emphasize the significance of the center patches, we offer a novel cropping approach based on the patch as a unit.

The Sparse pathway will pay greater attention to spatial features when we utilize the Sparse pathway to simulate the P-cells in the visual neural network. Based on the intuition that more actions occur in the center of the videos, we hope that the Sparse pathway will focus on the center of the sampled frame for learning. At the same time, the Dense pathway pays some attention to spatial features while extracting temporal features, resulting in improved spatiotemporal information fusion. Therefore, we propose a new crop method called Patch crop, which allows the network to extract patch-based local features more effectively.

We build SDTN on the idea of a frame that can be decomposed into many patches of 16 × 16 in size. Each frame is scaled to 224 × 224, which means that each frame will be divided into non-overlapping 196 patches. Then, on this basis, we will first resize the frame $(H \times W)$ to $((H + 2 \times \mathit{patch\_size}) \times (W + 2 \times \mathit{patch\_size}))$. Since the Transformer architecture cannot directly process frames, 2D CNNs are used to process the image into a feature map. The feature map is then split into 256 patches, with the middle 196 patches being sampled, as shown in Fig. 3.

These patches are linearly projected onto an embedding vector after being flattened into vectors:

$$z_{(p,t)}^{(0)} = Ex_{(p,t)} + e_{(p,t)}^{pos} \tag{1}$$

Here $x_{(p,t)} \in \mathbb{R}^{3P \times P}$ is the input vector, and the embedding vector $z_{(p,t)} \in \mathbb{R}^{D}$ is relevant to a learnable positional embedding vector $e_{(p,t)}^{pos}$, as well as the matrix E. The indices t, and p are the frame and patch index, respectively with $t = 1, \ldots, F$, and $p = 1, \ldots, N$. When using the Transformer model for action classification, we need to add a learnable classification token to the first position of the embedding sequence $z_{(0,0)}^{(0)} \in \mathbb{R}^{D}$.

$$q_{(p,t)}^{(l,a)} = W_{Q}^{(l,a)} LN(z_{(p,t)}^{(l-1)}) \in \mathbb{R}^{D_h} \tag{2}$$

$$k_{(p,t)}^{(l,a)} = W_{K}^{(l,a)} LN(z_{(p,t)}^{(l-1)}) \in \mathbb{R}^{D_h} \tag{3}$$

$$v_{(p,t)}^{(l,a)} = W_{V}^{(l,a)} LN(z_{(p,t)}^{(l-1)}) \in \mathbb{R}^{D_h} \tag{4}$$

Each pathway consists of L encoding blocks. At each block $\ell \in \{1, \ldots, L\}$, and head $a \in \{1, \ldots, \mathcal{A}\}$, we compute a query, key, and value vector for each patch based on the representation $z_{(p,t)}^{(\ell-1)}$ encoded of the preceding block. Where LN() represents LayerNorm [2]. The dimension of each self-attention head is set to $D_h = D/A$.

$$\alpha_{(p,t)}^{(\ell,a)} = \mathrm{SM}\left(\frac{q_{(p,t)}^{(\ell,a)\top}}{\sqrt{D_h}} \cdot \left[k_{(0,t)}^{(\ell,a)} \left\{ k_{(p',t')}^{(\ell,a)} \right\}_{\substack{p'=1,\ldots,N \\ t'=1,\ldots,F}} \right] \right) \tag{5}$$

Self-attention weights $\alpha_{(p,t)}^{(\ell,\alpha)} \in \mathbb{R}^{NF+F}$ are computed by dot-product. Where SM() represents the softmax activation function. Then, according to the research of STAM, global attention is applied to frames to realize action recognition. Finally, we utilize fusion method to combine the two networks' scores to derive the final prediction result. Formally, we employ the method of sampling and feeding twice, and separate it into two network pathways for modeling:

$$SDTN(F_s, F_d) = G(S_p(F_s), D_p(F_d)) \tag{6}$$

Here F_s stands for the Sparse frames sampled from the video, whereas F_d stands for the Dense frames sampled. To generate scores, F_s and F_d use their respective processing methods and then are fed into S_p (the Sparse pathway) and D_p (the Dense pathway). Based on the Sparse pathway and the Dense pathway scores, the fusion function G predicts the probability of each action class in the whole video. For G, we will utilize the commonly used max fusion.

4 Experiments

Implementation Details. SDTN comprises two parts: a Sparse pathway and a Dense pathway. In our experiments, each pathway strictly follows the hierarchical structure of the original STAM consisting of a spatial transformer and a temporal transformer. SDTN is one of the ViT family models and contains 12 Multi-head Self-Attention block layers, each with 12 self-attention heads. Among them, SDTN uses the imagenet-21K pretraining provided by [32]. The temporal transformer we employ only has 6-layers and 8-head self-attention since the time information is extracted at a deep level.

For inference, we use different time resolutions in the entire video to sample the frames twice and resize each frame so that the smaller dimension is 256. Then we random crop all sample frames of the video to a size of 256 × 256; we also apply random flip augmentation and auto-augment with Imagenet policy on all frames. After that, we execute patch crop to sample the central 196 patches. Finally, we use the same method for training and inference.

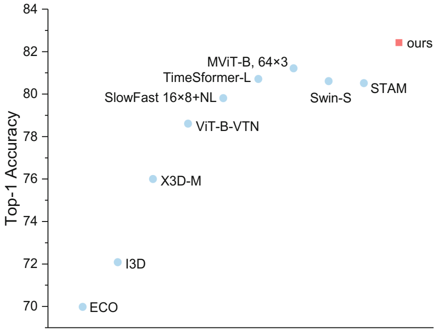

Fig. 4. Top-1 Accuracy on Kinetics-400 for the SDTN(16+64) *vs*. other networks. In terms of accuracy, SDTN is superior to previous architectures based on either CNNs, Transformer, or CNNs+Transformer.

Datasets. Kinetics-400 is a large-scale action recognition dataset, ∼240 k training videos and 20 k validation videos in 400 human action categories [19]. To further assess SDTN's performance on Kinetics-400, we build a Light kinetics-400 dataset based on Kinetics-400, allowing us to finetune our model according to the Kinetics-400 data format. The Light kinetics-400 dataset contains 400 human action categories. Especially, only five videos for each action category has been included to reduce the size of the dataset.

Training. We follow STAM and optimize from this basis. For frame alignment operation, we adjust the sampling rate and batch_size of each network to be the same.

5 Ablation Experiments

In this section, we provide the ablation experiments of SDTN on Kinetics-400 to evaluate the contributions of our two-stream Transformer architecture, frame alignment, and patch crop.

Table 1. Different frames combination

Sparse pathway	Dense pathway	Top-1 acc	Top-5 acc
16 frames	32 frames	81.93	95.52
32 frames	64 frames	81.68	95.34
16 frames	64 frames	**82.45**	**95.66**

The differences between various input frames are presented in Table 1. On the Kinetics-400 dataset, we can observe that our SDTN performs best when the Sparse pathway is 16 frames, and the Dense pathway input is 64 frames. However, when the Sparse pathway is 16 frames, and the Dense pathway is 64 frames, the input strategy is more accurate than when the Sparse path is 32 frames. This is because the Sparse pathway pays more attention to spatial information, and adding additional frames does not result in a significant increase in spatial information. Conversely, redundant frames increase computational cost and result in performance degradation.

On the Kinetics-400 dataset, we compare our approach with others. Table 2 indicates that, even when the number of GPUs is 4, each combination of input frames in SDTN can enhance accuracy and exceed the baseline. In Fig. 4, our model (82.45%) outperforms STAM's previous state-of-the-art by 1.95%.

Although SDTN does not reach SOTA accuracy, it shows promise as an effective Transformer model, that is, one that can explore the potential of a novel backbone based on conventional deep learning frameworks.

We design different patch crop experiments on SDTN(16+64). All Pathway means patch crop on the Dense pathway and the Sparse pathway simultaneously. As shown in Table 3, the effect of patch crop on the Sparse pathway is better than that on the Dense pathway, which is also constant with our intuition. The Sparse pathway pays extra attention to spatial features, making it more suitable for patch crop. Although the patch crop on the Dense pathway can exceed the baseline, the improvement is not significant for comparing with patch crop on the Sparse pathway. All Pathway performs well because in addition to extracting more spatial features, it can also better integrate spatiotemporal information. Actually, patch crop is more like extracting abstracts from a sentence.

Table 2. Comparision with other model on Kinetics-400

Model	Pretrain	Top-1 acc [%]	Top-5 acc [%]	Param [M]
I3D [7]	ImageNet-1K	72.1	90.3	25.0
X3D-M [10]	–	76.0	92.3	3.8
ip-CSN-152 [29]	–	77.8	92.8	32.8
ViT-B-VTN [23]	ImageNet-21K	78.6	93.7	114.0
X3D-XL [10]	–	79.1	93.9	11.0
SlowFast 16×8+NL [11]	–	79.8	93.9	59.9
MViT-B, 32×3 [9]		80.2	94.4	36.6
TimeSformer-L [3]	ImageNet-21K	80.7	94.7	121.4
MViT-B, 64×3 [9]	–	81.2	95.1	36.6
ViViT-L/16x2 320 [1]	ImageNet-21K	81.3	94.7	310.8
Swin-T [22]	ImageNet-1K	78.8	93.6	28.2
Swin-S [22]	ImageNet-1K	80.6	94.5	49.8
STAM(baseline) [26]	ImageNet-21K	80.5	–	96.0
SDTN(16+32)	ImageNet-21K	**81.9**	**95.5**	192.0
SDTN(32+64)	ImageNet-21K	**81.6**	**95.3**	192.0
SDTN(16+64)	ImageNet-21K	**82.4**	**95.6**	192.0

Table 3. Different patch crop pathway

Patch crop pathway	Top-1 acc	Top-5 acc
Dense pathway	82.11	95.50
Sparse pathway	82.31	95.65
All pathway	**82.45**	**95.66**

We compare the complete SDTN(16+64) with SDTN that only uses frame alignment or patch crop in Table 4. As can be seen from the table, both frame alignment and patch crop can enhance the accuracy of the network, surpassing some of the existing Transformer architectures.

Table 4. Frame alignment vs patch crop

Method	Top-1 acc	Top-5 acc
Only frame alignment	82.15	95.48
Only patch crop	81.68	95.54
Full SDTN	**82.45**	**95.66**

In Table 5, we employ several fusion strategies in order to achieve the full potential of SDTN(16+64). Among them, the weight fusion is inspired by Slow-Fast networks [11]. The P-cells and M-cells are interested in spatial and temporal

Table 5. Different fusion methods

Fusion method	Top-1 acc	Top-5 acc
AVG	79.51	94.23
Weight fusion	79.84	94.38
Max	**82.45**	**95.66**

features and account for ∼80% and ∼15–20% of the visual system, respectively. Consequently, we apply this ratio to SDTN, which combines the Dense pathway score of 20% and the Sparse pathway score of 80%.

The max fusion method is the most effective in experiments. That is because different actions have different requirements for temporal and spatial features. For example, recognizing some actions emphasizes spatial features, while others pay more attention to temporal features. The max fusion will select recognition results with more significant features and higher accuracy.

Table 6. Comparision on Light kinetics-400

Model	Pretrain	Top-1 acc	Top-5 acc
STAM(baseline)	ImageNet-21K	93.93	99.61
SDTN(16+32)	ImageNet-21K	**95.59**	**99.80**
SDTN(32+64)	ImageNet-21K	**95.40**	**99.80**
SDTN(16+64)	ImageNet-21K	**95.69**	**99.80**

Finally, as shown in Table 6, we evaluate the performance of baseline and SDTN on Light kinetics-400. It can be seen that when the size of the dataset decreases, the accuracy of the network rises substantially. On Light kinetics-400, the experimental results are consistent with the performance of SDTN on Kinetic-400, indicating that the dataset and SDTN are competent for the task of action recognition.

6 Conclusion

In this paper, a novel model Sparse Dense Transformer Network, a two-stream Transformer architecture, was proposed for action recognition. Patch crop was a new kind of cropping based on patch, which helps the network pays more attention to the patch in the center of the image. Frame alignment was adopted to assist the Dense pathway in selecting frames for input consistent with the Sparse pathway, improving accuracy while reducing the computational cost. The results of ablation experiments also show that the max fusion is the best fusion method for SDTN. Through extensive experiments in benchmarks, SDTN shows its superiority compared with the previous models, achieving 82.45% accuracy on the Kinetics-400. In the latter research, the extraction of local features for the patch will be considered into two-stream Transformer network for action recognition.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: a video vision transformer. arXiv preprint arXiv:2103.15691 (2021)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021)
4. Cao, W.P., et al.: An ensemble fuzziness-based online sequential learning approach and its application. In: International Conference on Knowledge Science, Engineering and Management (KSEM), pp. 255–267 (2021)
5. Cao, W., Xie, Z., Li, J., Xu, Z., Ming, Z., Wang, X.: Bidirectional stochastic configuration network for regression problems. Neural Netw. **140**, 237–246 (2021)
6. Cao, W., Yang, P., Ming, Z., Cai, S., Zhang, J.: An improved fuzziness based random vector functional link network for liver disease detection. In: 2020 IEEE 6th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), pp. 42–48 (2020)
7. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299–6308 (2017)
8. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fan, H., et al.: Multiscale vision transformers. arXiv preprint arXiv:2104.11227 (2021)
10. Feichtenhofer, C.: X3D: expanding architectures for efficient video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 203–213 (2020)
11. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 6202–6211 (2019)
12. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: action recognition by previewing audio. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10457–10467 (2020)
13. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 244–253 (2019)
14. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. arXiv preprint arXiv:2103.00112 (2021)
15. 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 (CVPR), pp. 770–778 (2016)

16. Hu, F., Lakdawala, S., Hao, Q., Qiu, M.: Low-power, intelligent sensor hardware interface for medical data preprocessing. IEEE Trans. Inf Technol. Biomed. **13**(4), 656–663 (2009)
17. Kahatapitiya, K., Ryoo, M.S.: Coarse-fine networks for temporal activity detection in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8385–8394 (2021)
18. Kalfaoglu, M.E., Kalkan, S., Alatan, A.A.: Late temporal modeling in 3D CNN architectures with BERT for action recognition. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12539, pp. 731–747. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-68238-5_48
19. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
20. Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., Sebe, N.: Spatio-temporal attention networks for action recognition and detection. IEEE Trans. Multimedia **22**(11), 2990–3001 (2020)
21. Li, Y., Song, Y., Jia, L., Gao, S., Li, Q., Qiu, M.: Intelligent fault diagnosis by fusing domain adversarial training and maximum mean discrepancy via ensemble learning. IEEE Trans. Industr. Inf. **17**(4), 2833–2841 (2020)
22. Liu, Z., et al.: Video Swin transformer. arXiv preprint arXiv:2106.13230 (2021)
23. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. arXiv preprint arXiv:2102.00719 (2021)
24. Qiu, H., Zheng, Q., Msahli, M., Memmi, G., Qiu, M., Lu, J.: Topological graph convolutional network-based urban traffic flow and density prediction. IEEE Trans. Intell. Transp. Syst. **22**(7), 4560–4569 (2020)
25. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5533–5541 (2017)
26. Sharir, G., Noy, A., Zelnik-Manor, L.: An image is worth 16 × 16 words, what is a video worth? arXiv preprint arXiv:2103.13915 (2021)
27. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199 (2014)
28. 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 (ICCV), pp. 4489–4497 (2015)
29. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5552–5561 (2019)
30. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6459 (2018)
31. Van Essen, D.C., Gallant, J.L.: Neural mechanisms of form and motion processing in the primate visual system. Neuron **13**(1), 1–10 (1994)
32. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS), pp. 5998–6008 (2017)
33. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2

34. Zhang, S., Guo, S., Huang, W., Scott, M.R., Wang, L.: V4D: 4d convolutional neural networks for video-level representation learning. arXiv preprint arXiv:2002.07442 (2020)
35. Zolfaghari, M., Singh, K., Brox, T.: ECO: efficient convolutional network for online video understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 713–730. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_43