Embedding Human Knowledge into Spatio-Temproal Attention Branch Network in Video Recognition via Temporal Attention

Saki Noguchi
noguchi@mprg.cs.chubu.ac.jp
Yuzhi Shi
shi@mprg.cs.chubu.ac.jp
Tsubasa Hirakawa
hirakawa@mprg.cs.chubu.ac.jp
Takayoshi Yamashita
takayoshi@isc.chubu.ac.jp
Hironobu Fujiyoshi
fujiyoshi@isc.chubu.ac.jp

Chubu University
1200 Matsumotocho
Kasugai, Aichi, Japan

Abstract

When recognizing objects or motions, humans can accurately judge the necessary areas for recognition. In contrast, recognition using a deep learning model is based on its training data and may fail to focus on the correct regions. In image recognition, it has been shown that visualizing the basis of decisions and embedding human knowledge into deep neural networks are effective in addressing this issue. However, in video recognition, there is no visualization method enabling us to embed human knowledge. We propose the spatio-temporal attention branch network (ST-ABN) for video recognition, which provides visual explanations for both spatial and temporal attentions. One of the features of the ST-ABN is that its attention output can be modified on the basis of human knowledge and used for recognition. However, since a video consists of a large number of frame images, modifying spatial attentions similar to image recognition is costly. Therefore, we manually modify temporal attentions to embed human knowledge into the ST-ABN. Experimental results with Something-Something v.2 indicate that the ST-ABN provides visual explanation for both spatial and temporal information and improves recognition performance. The results also indicate the effectiveness of embedding human knowledge into the ST-ABN and the positive changes in spatial attentions by modifying temporal attentions.

1 Introduction

Video recognition is a task for identifying actions performed in a video using multiple frame images. Many convolutional neural networks (CNNs) [1, 15] for video recognition, such as three-dimensional (3D) CNNs, have been proposed and achieve high recognition performance. For practical implementation of video recognition, it is important to ensure reliability

as well as recognition accuracy. In particular, deep learning models are expected to enable humans to understand the basis of their decisions. Visual explanation has been used to interpret the decision-making of CNNs by highlighting the gazing area during the inference process. Typical visual explanations include class activation mapping (CAM) [30], gradient-weighted class activation mapping (Grad-CAM) [22], and using the attention branch network (ABN) [8]. The ABN achieves high recognition accuracy by visualizing the gazing area as heat maps and weighting the feature maps. However, if the training data are biased, it cannot visualize at the correct regions, causing misrecognition.

To solve this problem, a method of embedding human knowledge into the model via attention map was proposed [21]. This method manually corrects the attention maps of misclassified images and fine-tunes the deep learning model. This enables the acquisition of an attention map closer to the human's gazing area. However, there is no method like this in the field of video recognition.

To address this issue, we propose the spatio-temporal attention branch network (ST-ABN) for visual explanation of video recognition that can embed human knowledge. The ST-ABN provides visual explanation and improves recognition performance by applying importance of spatial and temporal information to the recognition process. It also has an attention mechanism to weight attentions and applies them to recognition. Thus, we can change the parameters of the ST-ABN by using modified attentions.

The main contributions of this work are as follows. We propose the ST-ABN, a network model for providing visual explanation of video recognition that can embed human knowledge. To show the effectiveness of embedding human knowledge into the ST-ABN, we manually modify temporal attentions and fine-tune it. Finally, we evaluated the effectiveness of modifying temporal attentions on spatial attentions.

2 Related Work

Video Recognition Video recognition using deep learning is categorized into three types, 2D CNN-based, 3D CNN-based, and transformer-based. The 2D CNN-based methods [6, 23] typically use a two-stream network structure in which each frame and the optical flow of motion information are input to two separate CNNs. Each network extracts spatial features from video frames and temporal ones from an optical flow and is treated as independent. The 3D CNN-based methods use 3D convolution that extends a 2D convolution into temporal directions. They extract ST features by stacking multiple 3D convolution layers [13, 24]. Unlike 2D CNN-based methods, the extracted ST features take into account the interrelationship between spatial and temporal information. However, 3D CNN-based methods take into account only local relationships and are suitable for shorter videos. Transformer-based methods divide the frame images into patches [2, 3]. Patches at the same location in each frame are used to obtain temporal information, and those in the same frame are used to obtain spatial-information features. Transformer-based methods can learn global relationships faster than CNN-based ones.

Visual Explanation In image-classification tasks, many methods for visual explanation [8, 22, 30] have been proposed to analyze the basis of the decision by visualizing an attention map that shows the gazing area of the model. There are two methods for obtaining attention maps: one is using the gradient during backpropagation, and the other is using the response

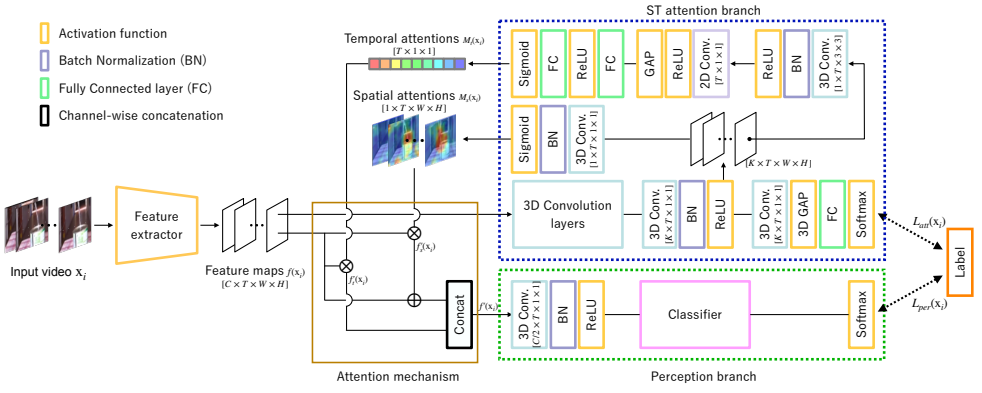

Figure 1: Detailed structure of ST-ABN. We divide the backbone network into the feature extractor and perception branch and add the ST attention branch between them.

from the network. Grad-CAM [22] is one way to use the gradient during backpropagation. It obtains attention maps by using a gradient of a particular class during the backpropagation process and can be applied to many pre-trained networks. One method of using the response from the network is CAM [30], which obtains attention maps from feature maps. Its attention maps are obtained using the feature maps in the convolution layer and weights at the fully connected layers at each channel. However, CAM degrades recognition accuracy due to the lack of spatial information caused by global average pooling (GAP) between the convolution and fully connected layers. To solve this problem, the attention branch network (ABN) was proposed [8]. It applies attention maps to the attention mechanism to improve recognition accuracy and visual explainability.

Recognition with a deep learning model is based on its training data and may fail to focus on the correct regions. In contrast, humans already have enough information through experience, and can accurately judge the necessary areas. Models using an attention mechanism weight the attention map as input during recognition, and inappropriate gazing areas induce misrecognition. To solve this problem, Mitsuhara *et al*. proposed a method of embedding human knowledge into CNNs [21]. With this method, attention maps are corrected manually and the deep learning model is fined-tuned to improve its performance.

In video-recognition tasks, there is a method for visual explanation [28] but none can embed human knowledge. However, in these tasks, inappropriate gazing areas occur in the spatial information that should be gazing at the object and the temporal information that should be focusing on the motion segment. In particular, temporal information is important for video recognition, and it also has been shown that using only appropriate frames that include motion improves recognition accuracy [27].

3 Proposed Method

The ST-ABN provides a visual explanation for spatial and temporal information. Since the ST-ABN uses an attention mechanism, we can embed human knowledge via both spatial and temporal attentions.

3.1 ST-ABN

As shown in Figure 1, The ST-ABN involves three modules: a feature extractor, ST attention branch, and perception branch. The feature extractor consists of multiple convolution layers and outputs feature maps from the inputs. We introduce an ST attention branch that outputs spatial and temporal attentions, which indicate the importance of spatial and temporal information to a network based on 3D CNNs. The perception branch inputs feature maps, which are weighted spatial and temporal attentions, by the attention mechanism and outputs the probability of each class.

3.1.1 Spatio-temporal Attention Branch

As shown in Figure 1, the ST attention branch generates spatial and temporal attentions that represent the importance of spatial and temporal information, respectively. It also outputs the classification results via 3D GAP. In the ST attention branch, the feature maps output from the feature extractor are first fed into the 3D convolution layers consisting of multiple residual blocks, which have the same structure as the perception branch. We set the stride of the convolution layer at the first residual block to 1 to maintain the resolution of the feature maps. The feature maps from the 3D convolution layers are then input to another 3D convolution layer of $K\times T\times 1\times 1$, where K indicates the number of classes and T indicates the number of frames. Thus, the size of feature maps becomes $K\times T\times W\times H$. These feature maps are then input to the $K\times T\times 1\times 1$ 3D convolution layer, 3D GAP, and a softmax function to obtain the classification probabilities for each class. Another study of the ST attention branch is in the supplementary material.

Spatial Attentions We generate spatial attentions from the above-mentioned $K\times T\times W\times H$ feature maps. We apply a $1\times T\times 1\times 1$ 3D convolution layer for the $K\times T\times W\times H$ feature maps and obtain a single $1\times T\times W\times H$ feature map for each frame. This means that we aggregate the K feature maps with respect to each video frame into a single feature map. We can then obtain the spatial attentions M_s for each frame by applying a sigmoid function.

Temporal Attentions Similar to spatial attentions, we generate temporal attentions from the $K\times T\times W\times H$ feature maps. These feature maps are first aggregated into a single $1\times T\times W\times H$ feature map with respect to each frame by applying a $1\times T\times 3\times 3$ 3D convolution layer. The channel dimensions of the $1\times T\times W\times H$ feature maps are then reduced and transformed into $T\times W\times H$ feature maps. The $T\times W\times H$ feature maps are further input to a $T\times 1\times 1$ 2D convolution layer, and the mean value of each feature map in the spatial direction is calculated by GAP. Finally, the temporal attentions M_t is generated via the fully connected layer, rectified linear unit (ReLU), and a sigmoid function. We use a simple gating mechanism that uses a sigmoid function such as squeeze-and-excitation networks (SENet) [12]. This enables the ST-ABN to emphasize multiple frames instead of only a single frame.

3.1.2 Attention Mechanism

The spatial and temporal attentions acquired from the ST attention branch are further used as an attention mechanism for weighting feature maps. Let $\mathbf{x}_i$ be the i-th video in a dataset and $f(\mathbf{x}_i)$ be the corresponding feature maps obtained from the feature extractor. The weighted feature maps $f'_s(\mathbf{x}_i)$ by spatial attentions $M_s(\mathbf{x}_i)$ is defined as

$$f'_s(\mathbf{x}_i) = (1+M_s(\mathbf{x}_i))\cdot f(\mathbf{x}_i). \tag{1}$$

For spatial attentions, we apply a residual mechanism [25] and add the unweighted feature maps to the weighted feature maps. This can suppress the disappearance of the feature maps, and the attention maps can be efficiently reflected in the recognition. The attention mechanism with $M_t(\mathbf{x}_i)$ calculates the weighted feature maps $f'_t(\mathbf{x}_i)$ as

$$f'_t(\mathbf{x}_i) = M_t(\mathbf{x}_i) \cdot f(\mathbf{x}_i). \tag{2}$$

For the temporal attentions, we apply simple weighting and do not use a residual attention mechanism.

These $f'_s(\mathbf{x}_i)$ and $f'_t(\mathbf{x}_i)$ are combined in the channel direction by

$$f'(\mathbf{x}_i) = \mathrm{conv}_\theta\left(\mathrm{concat}[f'_s(\mathbf{x}_i), f'_t(\mathbf{x}_i)]\right), \tag{3}$$

where $f'(\mathbf{x}_i)$ denotes the concatenated feature maps. The number of channels is doubled because the two feature maps are channel-wise concatenated.

3.1.3 Training

In this section, we explain the implementation of the ST-ABN. The ST-ABN is constructed by dividing a backbone network into a feature extractor and perception branch and adding an ST attention branch between them. As a result, it can be easily introduced into various baseline models (e.g., C3D, 3D ResNet). In this study, we used 3D ResNet, which is a temporally inflated version of ResNet [10], as the backbone network. Specifically, ST-ABN is constructed using 3D ResNet on the basis of the slow pathway of SlowFast networks [7]. The spatial dimension of the input is 224×224, and the input data size is $C\times T\times W\times H$. To suppress overfitting, we apply a dropout [11] of 0.5 to both the ST attention branch and the perception branch.

The loss function of ST-ABN $\mathcal{L}(\mathbf{x}_i)$ is calculated as

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i), \tag{4}$$

where $\mathcal{L}_{att}(\mathbf{x}_i)$ and $\mathcal{L}_{per}(\mathbf{x}_i)$ denotes the training loss at the ST attention branch and perception branch, respectively. The $\mathcal{L}_{att}(\mathbf{x}_i)$ and $\mathcal{L}_{per}(\mathbf{x}_i)$ can be calculated by the softmax function and cross-entropy error. The loss function of the ST-ABN is trained in an end-to-end manner.

3.2 Embedding Human Knowledge via Temporal Attentions

The attentions obtained from the ST attention branch sometimes indicate inappropriate gazing areas, similar to other visual explanation methods. However, ST-ABN can embed human knowledge thanks to its attention mechanism. Therefore, we manually modified the attentions and fine-tuned the ST-ABN. In this paper, we only modified the temporal attentions, as modifying the attentions is highly costly, as shown in the supplementary material.

3.2.1 Fine-tuning with Modified Temporal Attentions

Temporal attentions are modified in three steps, as shown in Figure 2.

Step 1 Train the ST-ABN and collect the temporal attentions of misclassified videos.
Step 2 Manually modify the temporal attentions collected in step 1.
Step 3 Fine-tune the branches of ST-ABN with the modified temporal attentions in step 2.

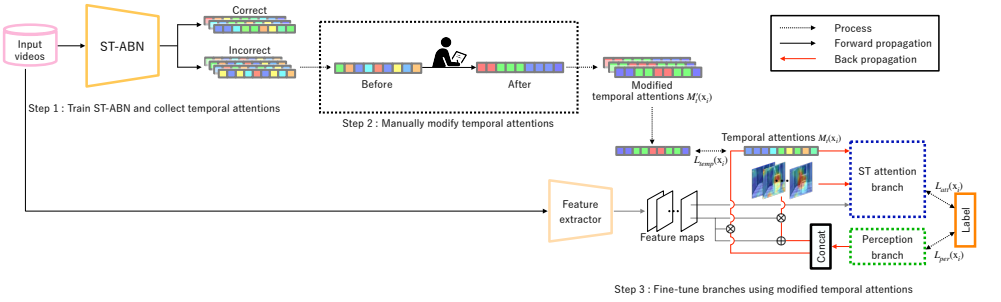

Figure 2: Flow of embedding human knowledge into the ST-ABN via temporal attentions.

An example of temporal-attention modification is in the supplementary material.

After modifying the temporal attentions, the ST-ABN embeds human knowledge by being fine-tuned with these attentions. The loss function of fine-tuning $\mathcal{L}(\mathbf{x}_i)$ is defined as

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i) + \mathcal{L}_{temp}(\mathbf{x}_i). \tag{5}$$

In this process, we add $\mathcal{L}_{temp}(\mathbf{x}_i)$ to that of the ST-ABN calculated with $\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i)$. By adding $\mathcal{L}_{temp}(\mathbf{x}_i)$, temporal attentions obtained from the ST-ABN become closer to the modified attentions. As for the loss of the temporal attentions $\mathcal{L}_{temp}(\mathbf{x}_i)$, we use the mean squared error of these temporal attentions. We denote the output temporal attentions from the ST-ABN and modified temporal attentions as $M_t(\mathbf{x}_i)$ and $M'_t(\mathbf{x}_i)$, respectively. The $\mathcal{L}_{temp}(\mathbf{x}_i)$ is formulated as

$$L_{temp}(\mathbf{x}_i) = \gamma_t \frac{1}{n} \sum_{j=1}^{n} \left(\{M'_t(\mathbf{x}_i)\}_j - \{M_t(\mathbf{x}_i)\}_j \right)^2, \tag{6}$$

where n is the number of input frames, and γ_t is a scale factor.

Thus, we can embed human knowledge into the ST-ABN via modified temporal attentions by fine-tuning the ST-ABN. During the fine-tuning, the ST-ABN optimizes its ST attention and perception branches. The feature extractor, which extracts the feature maps from an input video, is not updated during the fine-tuning process.

4 Experiments

We evaluated the effectiveness of embedding human knowledge into the ST-ABN using Something-Something v.2, which is a benchmark for action recognition. We first compared the recognition accuracy of the ST-ABN, with those of conventional models. We also qualitatively and quantitatively evaluated the explainability of spatial and temporal attentions.

4.1 Experiment Details

Something-Something v.2 [8] is used as a benchmark for large-scale action recognition with 220,847 videos and recognizes 174 basic actions of a person handling everyday objects. The length of the videos ranges from 2 to 6 seconds. To embed human knowledge into the ST-ABN via temporal attentions, we modified eight actions in which recognition accuracy in

Table 1: Performance evaluation (top-1 and top-5 accuracy) of each model on Something-Something v.2. Accuracy of 3D ResNet-50 and 3D ResNet-101 improved by introducing the ST-ABN and on par with other conventional models.

Model	Backbone	Frames	Top-1	Top-5
TSN [26]	BN-Inception	8	27.8	57.6
TRN Multiscale [31]	BN-Inception	8	48.8	77.6
TRN Two-stream [31]	BN-Inception	16	55.5	83.1
CPNet [18]	ResNet-34	24	57.7	84.0
TSM [17]	ResNet-50	8	59.1	85.6
TSM [17]	ResNet-50	16	63.4	88.5
STM [14]	ResNet-50	8	62.3	88.8
STM [14]	ResNet-50	16	64.2	89.8
GST [19]	ResNet-50	16	62.6	87.9
ABM [33]	ResNet-50	16×3	61.3	–
DFB-Net [20]	ResNet-152	16	57.7	84.0
bLVNet-TAM [5]	bLResNet-101	32×2	65.2	90.3
SmallBig$_{En}$ [16]	ResNet-50	24×2×3	64.5	89.1
PEM [29]	ResNet-50	16×2	65.0	–
Zhou *et al.* [32]	3D DenseNet-121	16	62.9	88.0
3D ResNet-50 (Our baseline)	–	32	51.4	80.1
ST-ABN (Ours)	3D ResNet-50	32	**58.6**	**85.5**
3D ResNet-50 (Our baseline)	–	32×2	63.8	89.2
ST-ABN (Ours)	3D ResNet-50	32×2	**64.1**	**89.6**
3D ResNet-101 (Our baseline)	–	32	57.7	82.8
ST-ABN (Ours)	3D ResNet-101	32	**58.0**	**83.2**
3D ResNet-101 (Our baseline)	–	32×2	65.3	90.1
ST-ABN (Ours)	3D ResNet-101	32×2	**65.8**	**90.4**

both training and evaluation data were low. Seventy-four people manually modified 2396 videos and used them in the fine-tuning process. The backbone networks of the ST-ABN were 3D ResNet-50 and 3D ResNet-101. Regarding for the number of frames to be input, we compared the recognition accuracy between the case of inputting 32 frames selected at random and that of inputting two sets of 32 frames in which the input video was randomly divided into two segments. We optimized the networks by stochastic gradient descent with momentum and set a momentum and weight decay of 0.9 and 0.0005, respectively. We used over 8 GPUs, and each GPU had a batch size of 8, resulting in a mini-batch of 64 in total. Our models were initialized using pre-trained models on ImageNet [4]. All models started training at a learning rate of 0.01, and the learning rate was multiplied by $1/10$ after the saturation of the validation loss. During fine-tuning, we stared training at a learning rate of 0.0001 and the scale factor γ_t of 10. The other experimental details were the same as those for training the ST-ABN.

4.2 Experimental Results

We quantitatively evaluated the ST-ABN and conventional models, and its effectiveness in embedding human knowledge.

Comparison with Conventional Models We compared the performances of various conventional models with that of the ST-ABN. Table 1 shows the top-1 and top-5 accuracies of the comparison. It shows that introducing the ST-ABN into the backbone network, 3D

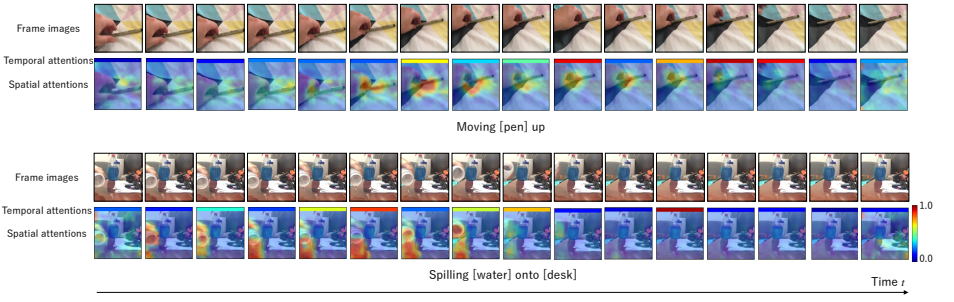

Figure 3: Visualization results of spatial and temporal attentions of two videos. From top to bottom, each figure shows input video frames and corresponding temporal attentions, spatial attentions, and action class.

ResNet, improved recognition accuracy and performed equally to or greater than the other models. This indicates that recognition accuracy can be improved by applying spatial and temporal attentions to the recognition process through the attention mechanism.

Comparison with Embedding Human Knowledge We compared the performance of ST-ABN before and after fine-tuning. The backbone network and input frames used in experiment were 3D ResNet-50 and 32, respectively. We selected eight target actions that had less than 50% classification accuracy on both the training and evaluation data, and modified their temporal attentions. The results are listed in Table 2. By being fine-tuned the ST-ABN, the actions with modified temporal attentions improved by 5.8%, and the actions without such a process also improved by 1.9%. This result indicates that embedding human knowledge is effective in improving recognition accuracy. Furthermore, modifying temporal attentions in some actions improves the performance of other action classes that are similar to be modified.

Table 2: Performance of top-1 accuracy before and after fine-tuning. The 'Modified' column shows the accuracy for 8 action classes with modified temporal attentions, while the 'Other' shows the accuracy for the remaining 166 action classes.

	Modified	Other	All
Before	20.5	59.8	58.6
After	**26.3**	**61.7**	**60.7**

4.3 Evaluation of Attentions

We visualized spatial and temporal attentions through qualitative evaluation.

ST-ABN Figure 3 shows examples of visualized spatial and temporal attentions. For spatial attentions, we visualized the attention maps of each frame as heat maps. The color bar corresponding to each frame is a color representation of the weight of a temporal attention.

From the visualization results of spatial attentions, they strongly highlight the regions handling any object and weakly highlight the other regions. From the results of temporal attentions, large weight outputs correspond to frames with the motion representing the class of action recognition. These results indicate that the ST-ABN can provide visual explanation that takes into account both spatial and temporal information simultaneously.

ST-ABN with Embedded Human Knowledge Figure 4 shows the visualization results of

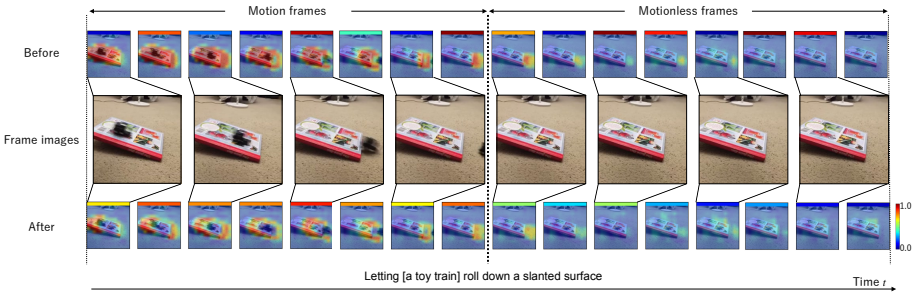

Figure 4: Visualization results from spatial and temporal attentions before and after fine-tuning. The visualization result before is on top, and the bottom is after. Frame images are shown in the middle of these results.

spatial and temporal attentions before and after fine-tuning the ST-ABN. Before fine-tuning, the changes in the color bar of temporal attentions are large regardless of the presence or absence of motion. In contrast, after fine-tuning, the motion was correctly recognized, and the changes of it were gradual. This indicates that the introduction of human knowledge into the ST-ABN enabled the acquisition of the appropriate attention focused on the motion segment. As for spatial attentions, before fine-tuning, the ST-ABN focuses on moving objects, whereas after fine-tuning, it focuses on the changed area between frames. Since it sometimes treats the same objects in the different action classes of video, action recognition using Something-Something v.2 needs to focus on the area changed by the motion, rather than the object. This means we could obtain better spatial attentions.

4.4 Effect to Spatial Attentions

We inverted spatial attentions to quantitatively evaluate the effect of modifying temporal attentions on it. We compared the recognition accuracy of the ST-ABN with and without inverting spatial attentions to confirm the effectiveness of spatial information for each ST-ABN version. The spatial attentions are inverted by

$$M_{s\ \mathrm{invert}}(\mathbf{x}_i) = 1 - M_s(\mathbf{x}_i), \tag{7}$$

where $M_s(\mathbf{x}_i)$ represents a spatial attention, and $M_{s\ \mathrm{invert}}(\mathbf{x}_i)$ represents an inverted spatial attention.

Table 3 lists the results of comparing the recognition accuracy of the ST-ABN by reversal of the spatial attentions. The top-1 and top-5 accuracies of the ST-ABN before fine-tuning decreased 31.5 and 32.8%, respectively, by inverting the spatial attentions. In contrast, those accuracies after fine-tuning decreased 40.6% and 43.6%, which is a higher rate than before the ST-ABN is fine-tuned. This indicates that modifying temporal attentions and fine-tuning the ST-ABN with them has a positive effect on spatial attentions.

Table 3: Accuracy of before and after fine-tuning the ST-ABN. A checkmark indicates spatial attentions were inverted. When decreasing rate of inverting attentions is greater, the ST-ABN obtained more effective attentions.

(a) Before

invert	Top-1	Top-5
	58.6	85.5
✓	27.1	52.7

(b)After

invert	Top-1	Top-5
	60.7	86.9
✓	20.1	43.4

5 Conclusion

We proposed the spatio-temporal attention branch network (ST-ABN) for providing a visual explanation of video recognition and embedding human knowledge via temporal attentions. The ST-ABN acquires the importance of spatial and temporal information, which can be applied to the attention mechanism to improve visual explanation and recognition performance. Those are further improved by embedding human knowledge. Furthermore, we found that modifying temporal attentions enables spatial attentions to acquire a more effective gazing area. Our future work is to extend the ST-ABN for a transformer-based method.

Acknowledgment

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Krizhevsky Alex, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[5] Quanfu Fan, Chun-Fu (Richard) Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3d779cae2d46cf6a8a99a35ba4167977-Paper.pdf.

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[8] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *2019 IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019.

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision*, volume 1, page 5, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[14] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *2019 IEEE International Conference on Computer Vision*, pages 2000–2009, 2019.

[15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[16] Xianhang Li, Yali Wang, Zhipeng Zhou, and Yu Qiao. Smallbignet: Integrating core and contextual views for video classification. In *2020 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1092–1101, 2020.

[17] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *2019 IEEE International Conference on Computer Vision*, pages 7083–7093, 2019.

[18] Xingyu Liu, Joon-Young Lee, and Hailin Jin. Learning video representations from correspondence proposals. In *2019 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4273–4281, 2019.

[19] Chenxu Luo and Alan L Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *2019 IEEE International Conference on Computer Vision*, pages 5512–5521, 2019.

[20] Brais Martinez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe. Action recognition with spatial-temporal discriminative filter banks. In *2019 IEEE International Conference on Computer Vision*, pages 5482–5491, 2019.

[21] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge into deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 2019.

[22] Selvaraju Ramprasaath, R., Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[23] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[25] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[26] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[27] Yang Wang and Minh Hoai. Improving human action recognition by non-action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] Yang Wang, Vinh Tran, Gedas Bertasius, Lorenzo Torresani, and Minh Hoai Nguyen. Attentive action and context factorization. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0187.pdf.

[29] Junwu Weng, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Xudong Jiang, and Junsong Yuan. Temporal distinct representation learning for action recognition. In *European Conference on Computer Vision*, pages 363–378. Springer, 2020.

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[31] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision*, pages 803–818, 2018.

[32] Yizhou Zhou, Xiaoyan Sun, Chong Luo, Zheng-Jun Zha, and Wenjun Zeng. Spatiotemporal fusion in 3d cnns: A probabilistic view. In *2020 IEEE Conference on Computer Vision and Pattern Recognition*, pages 9829–9838, 2020.

[33] Xinqi Zhu, Chang Xu, Langwen Hui, Cewu Lu, and Dacheng Tao. Approximated bilinear modules for temporal modeling. In *2019 IEEE International Conference on Computer Vision*, pages 3494–3503, 2019.