Embedding Human Knowledge into Spatio-Temproal Attention Branch Network in Video Recognition via Temporal Attention (Supplementary material)

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

In this supplementary document, we report the structure of the ST attention branch, the modification details of temporal attentions, and additional experiments for modifying attentions.

1 Study on ST Attention Branch

1.1 Structure Details

Figure 1 shows a structure of the other ST attention branch. This model is the same as the one in the main paper until creating $K \times T \times 1 \times 1$ feature maps. However, instead of convolving spatial and temporal attentions in parallel, temporal attentions are obtained in this structure using dimensionality compression with global average pooling after the spatial attentions are obtained.

1.2 Experiments

We quantitatively and qualitatively evaluated the ST-ABN with a series structure as shown in Figure 1. The backbone network of the ST-ABN was 3D ResNet-50, and 32 frames were inputted. The other experimental details were the same as those for training the ST-ABN in the main paper.

Quantitative Evaluation We compared the performance of our baseline model with that of the ST-ABN. We used two types of ST-ABN: the supplementary structure, in which

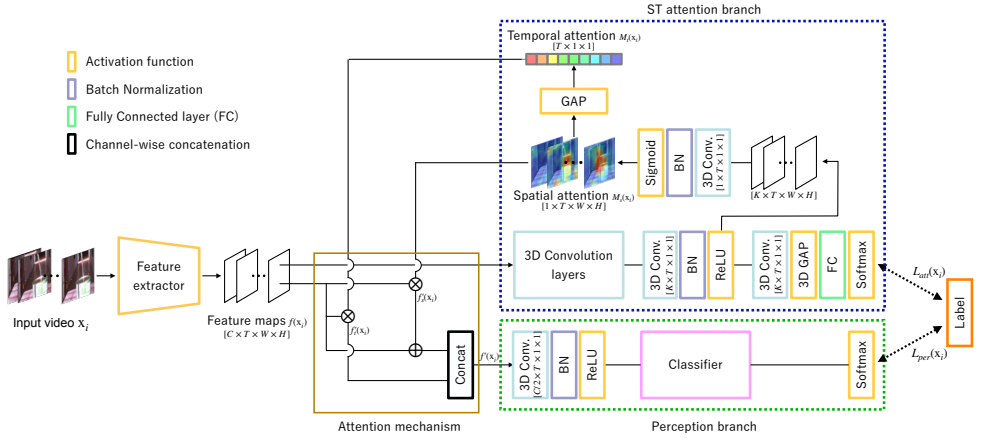


Figure 1: Detailed structure of ST-ABN, in which spatio-temporal information is calculated in a series.

Table 1: Performance of top-1 and top-5 accuracy of each model.

	Top-1 Acc.	Top-5 Acc.
3D ResNet-50 (Our baseline)	51.4	80.1
3D ResNet-50 + ST-ABN (Supplementary)	58.0	85.2
3D ResNet-50 + ST-ABN (Ours)	**58.6**	**85.5**

spatio-temporal information is calculated in a series, and the main structure, in which spatio-temporal information is calculated in parallel. In the table 1, both ST-ABN structures had higher recognition accuracies than those of the baseline model, but the ST-ABN structure described in the main paper was slightly higher. This indicates that the information necessary for recognition could be obtained by convolving spatial and temporal information in the same place. However, convoluting them separately and obtaining the features specialized for each kind of information are better options.

Qualitative Evaluation Figure 2 shows the visualization results of the spatial and temporal attentions of the ST-ABN with two types of structures. The results of the ST-ABN were that spatial attentions were able to focus on the moving object in both structures. However, the temporal attentions of the ST-ABN with the supplementary structure were almost all yellowish-green, making it impossible to determine which frames are important for recognition. This indicates that we need to design separate networks for spatial and temporal attentions.

2 Modification Details of Temporal Attentions

The temporal attentions were modified manually using a tool that we made, as shown in Figure 3. This tool enables us to edit temporal attentions easily—only by double clicking on the frame image. Temporal attentions were visualized on the top of each frame image as a color bar. When we modified the temporal attentions, the frames that were least important and not needed for recognition were given blue color bars with a value of 0, the frames with

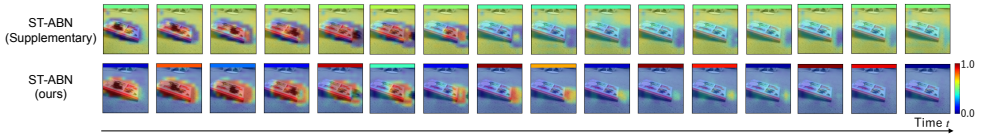
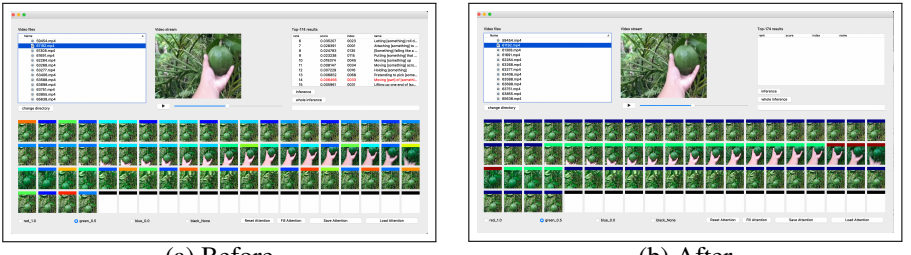


Figure 2: Visualization results from the spatial and temporal attentions of each structure. This is an example of a video in which a toy train is rolling down on a slanted surface. The temporal attentions are at the top of the frame images, and the spatial attentions are on the frame images.



(a) Before (b) After

Figure 3: Tool for modifying temporal attentions. The attention score was set with a radio button at the bottom of the tool, and temporal attentions were modified by double clicking each frame image. (a) Before modifying temporal attentions. (b) After modifying temporal attentions.

a necessary motion were given green ones with a value of 0.5, and the frames with a particularly important motion were given red ones with a value of 1.0. An example of temporal attention modification is shown in Figure 4. These temporal attentions were modified so that the frames with motion were highlighted.

3 Additional Study on Modification

Spatial and temporal attentions obtained from the ST-ABN enable us to embed human knowledge by modifying them in manual operation. Therefore, we investigated which attentions are more appropriate to modify from their cost and effectiveness of modification.

3.1 Modification Cost

We compared the modification cost by measuring the time required for modifying these attentions. Temporal attentions were modified for approximately 30 videos per hour because we just needed to highlight the important frame images. In contrast, the spatial attentions could only modify approximately five videos per hour because we had to highlight the object for each frame image. Note that the time taken to modify the spatial attentions was highly dependent on the number of frame images in the video, indicating that the modification cost of temporal attentions was much lower than that of spatial attentions.

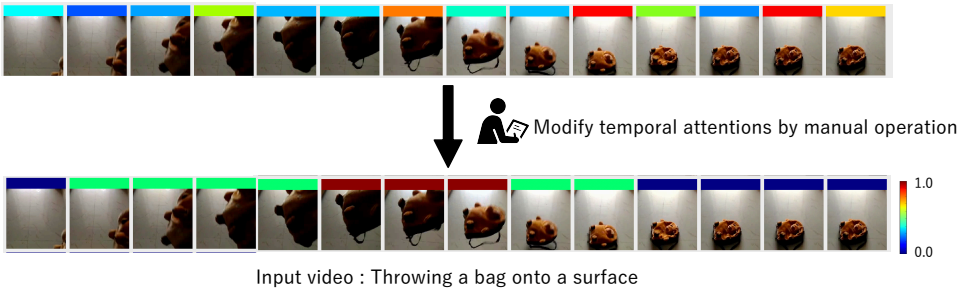


Figure 4: An example of modifying temporal attentions. We categorized each frame image into three types to modify. The temporal attentions are shown as a color bar on top of each frame image.

3.2 Effectiveness of Modification

We evaluated the effectiveness of modification by comparing the changes in accuracy before and after fine-tuning the ST-ABN with each attention.

3.2.1 Experiment Details

In the fine-tuning process, we added $\mathcal{L}_{temp}(\mathbf{x}_i)$ or $\mathcal{L}_{spat}(\mathbf{x}_i)$ to the loss function of the ST-ABN calculated by $\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i)$. When the ST-ABN is fine-tuned with modified temporal attentions, the loss function of the fine-tuning $\mathcal{L}(\mathbf{x}_i)$ can be defined as

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i) + \mathcal{L}_{temp}(\mathbf{x}_i), \tag{1}$$

where $\mathcal{L}_{temp}(\mathbf{x}_i)$ is the same as that in the main paper. When it is fine-tuned with modified spatial attentions, the $\mathcal{L}(\mathbf{x}_i)$ can be defined as

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{att}(\mathbf{x}_i) + \mathcal{L}_{per}(\mathbf{x}_i) + \mathcal{L}_{spat}(\mathbf{x}_i). \tag{2}$$

As for the loss function of the spatial attentions $\mathcal{L}_{spat}(\mathbf{x}_i)$, we used the mean squared error of modified and obtained attentions. We denote the output spatial attentions from the ST-ABN and modified spatial attentions as $M_s(\mathbf{x}_i)$ and $M'_s(\mathbf{x}_i)$, respectively. The $\mathcal{L}_{spat}(\mathbf{x}_i)$ are formulated as

$$L_{spat}(\mathbf{x}_i) = \gamma_s \frac{1}{n} \sum_{j=1}^{n} \left(\{M'_s(\mathbf{x}_i)\}_j - \{M_s(\mathbf{x}_i)\}_j \right)^2, \tag{3}$$

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

We manually modified 2396 temporal attentions and 464 spatial attentions to fine-tune the ST-ABN. The scale factor γ_t and γ_s were 10 and 100, respectively. The other experimental details were the same as those for fine-tuning the ST-ABN in the main paper.

3.2.2 Experimental Results

We qualitatively and quantitatively evaluated each model.

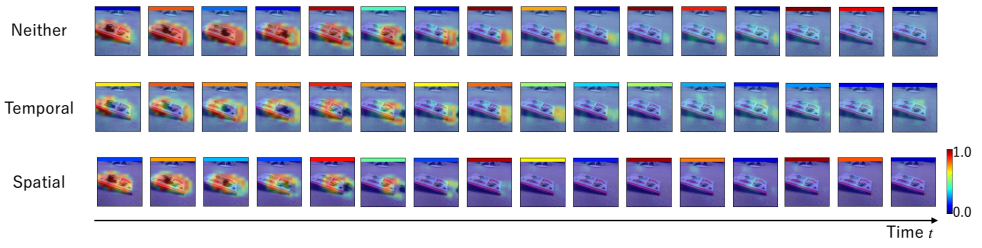


Figure 5: Visualization results of spatial and temporal attentions. The temporal attentions are at the top of the frame images, and the spatial attentions are on the frame images. Attention maps with "Neither" are the results of the ST-ABN before fine-tuning. The other attention maps with "Temporal" and "Spatial" are the results of the ST-ABN after fine-tuning via temporal and spatial attentions, respectively.

Quantitative Evaluation The results of fine-tuning with each attention are shown in Table 2. The recognition accuracy was improved by fine-tuning the ST-ABN by comparing the top line and the others. However, the recognition accuracy after fine-tuning was almost equal for all of them, indicating that the type of attentions to be modified did not affect the recognition accuracy.

Table 2: Accuracy of before and after embedding human knowledge by fine-tuning the ST-ABN. The top line without a checkmark shows the accuracy of ST-ABN before fine-tuning, and the other lines show the accuracy of ST-ABN after fine-tuning. A checkmark indicates its attentions were modified.

Spatial	Temporal	Top-1	Top-5
		58.6	85.5
	✓	**60.7**	**86.9**
✓		60.0	86.5

Qualitative Evaluation Figure 5 shows examples of visualized spatial and temporal attentions. When we fine-tuned the ST-ABN with modified temporal attentions, they were improved to focus on the frames in motion. Furthermore, spatial attentions also changed to highlight the area changed by the motion. In contrast, when the ST-ABN was fine-tuned with modified spatial attentions, they could focus on the local area, but the temporal attentions were almost the same as those before fine-tuning. This indicates that embedding human knowledge into the ST-ABN via temporal attentions is more effective to improve its performance.