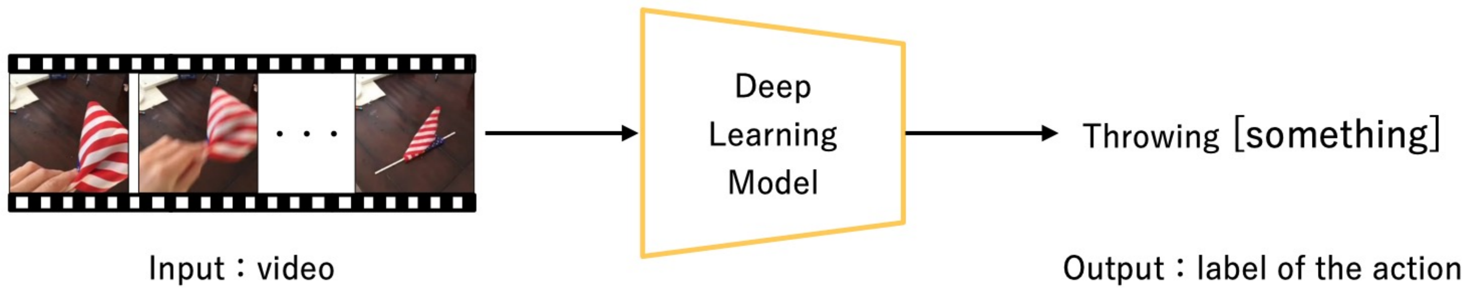
Embedding Human Knowledge into Spatio-Temproal Attention Branch Network in Video Recognition via Temporal Attention

Saki Noguchi, Yuzhi Shi, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi
Chubu University

MPRG MACHINE PERCEPTION AND ROBOTICS GROUP
BMVC 2023
ST-ABN
Attention Editor

Background and Motivation

- Video recognition by deep learning
  - A task for identifying actions performed in a video using multiple frame images
  - Recognize action using both spatial and temporal information



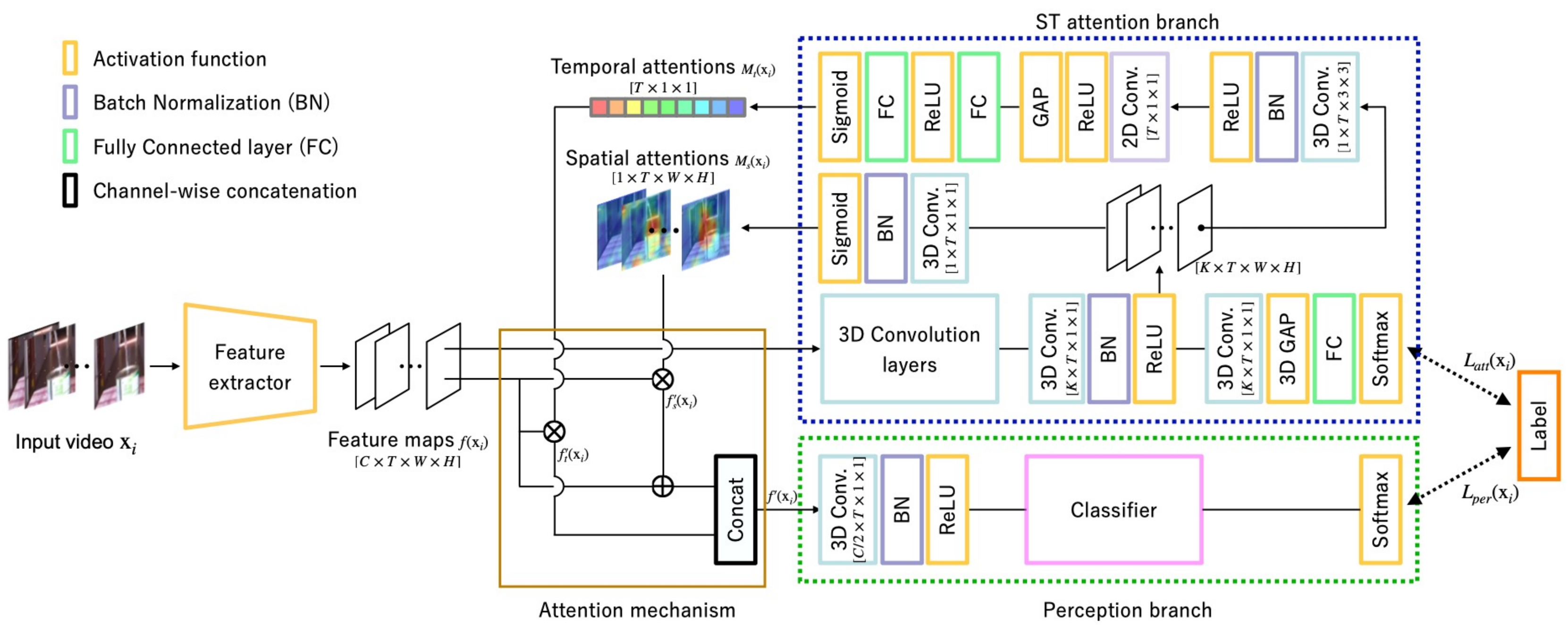
Problem : The basis of the model's decisions is unclear

Approach

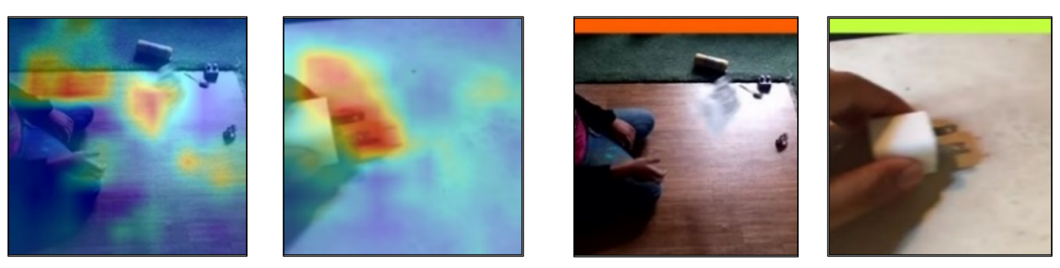
- ST-ABN
  - Visual explanation considering spatial and temporal information
- Embedding human knowledge into the ST-ABN
  - Improvement of recognition accuracy and visual explainability

Spatio-Temporal Attention Branch Network (ST-ABN)

- A network that takes into account important spatio-temporal information
  - ST attention branch : provide visual explanation for spatial and temporal attentions
    - Spatial attentions : Visualize the gazing area for each frame
    - Temporal attentions : Visualize the importance of each frame
  - Attention mechanism : Weight two attentions on the feature maps
- We can embed human knowledge via both spatial and temporal attentions



$$L(\mathrm{x}_i) = L_{att}(\mathrm{x}_i) + L_{per}(\mathrm{x}_i)$$

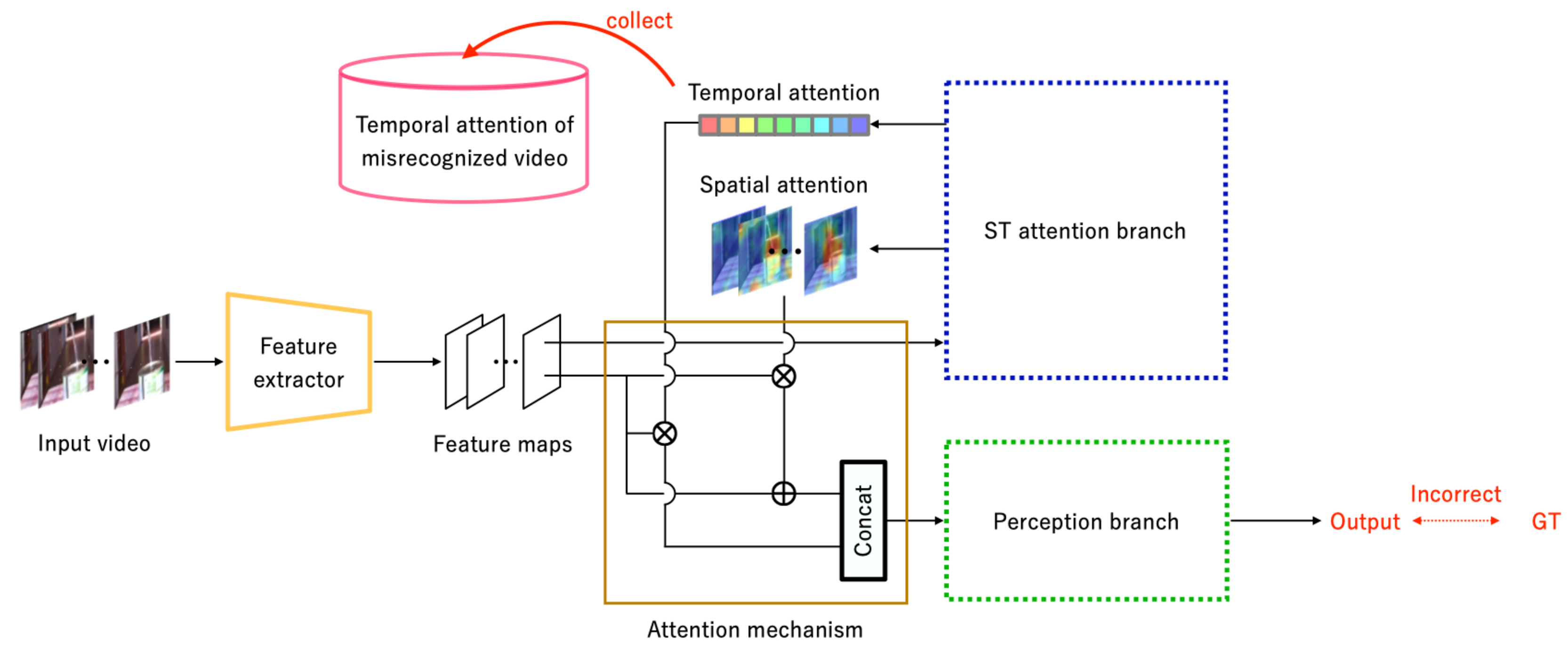


Spatial attentions Temporal attentions

Embedding human knowledge into the ST-ABN

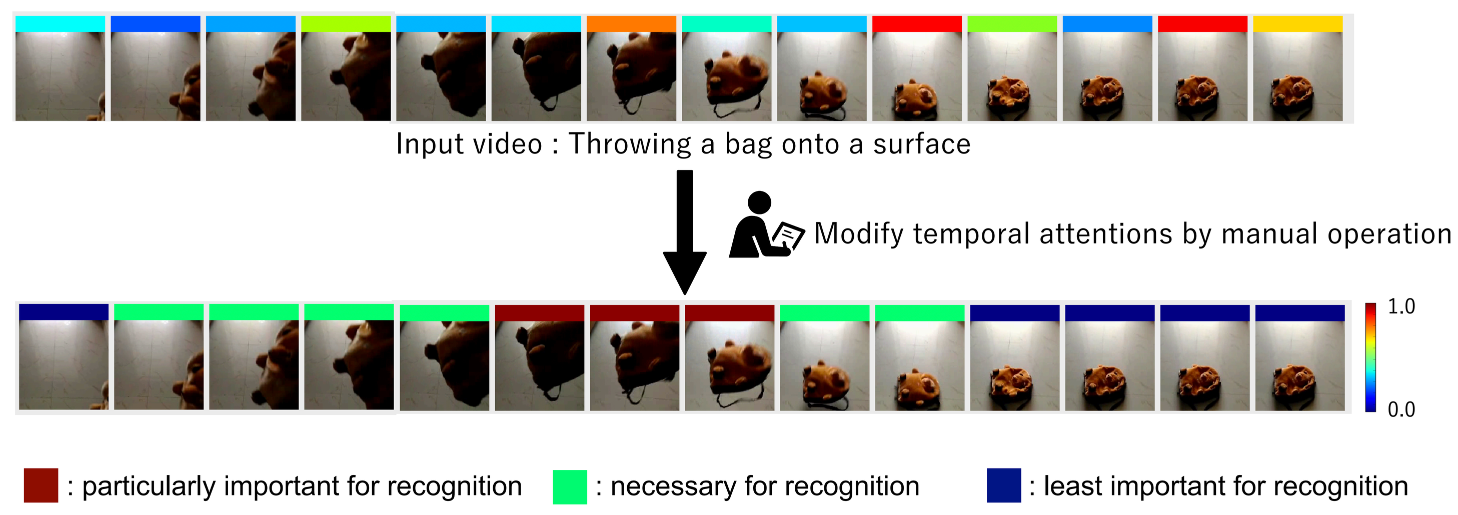
Step 1. Collecting temporal attentions

- Train the ST-ABN and collect temporal attentions
  - Evaluate with training samples and select misclassified videos



Step 2. Temporal attentions modification

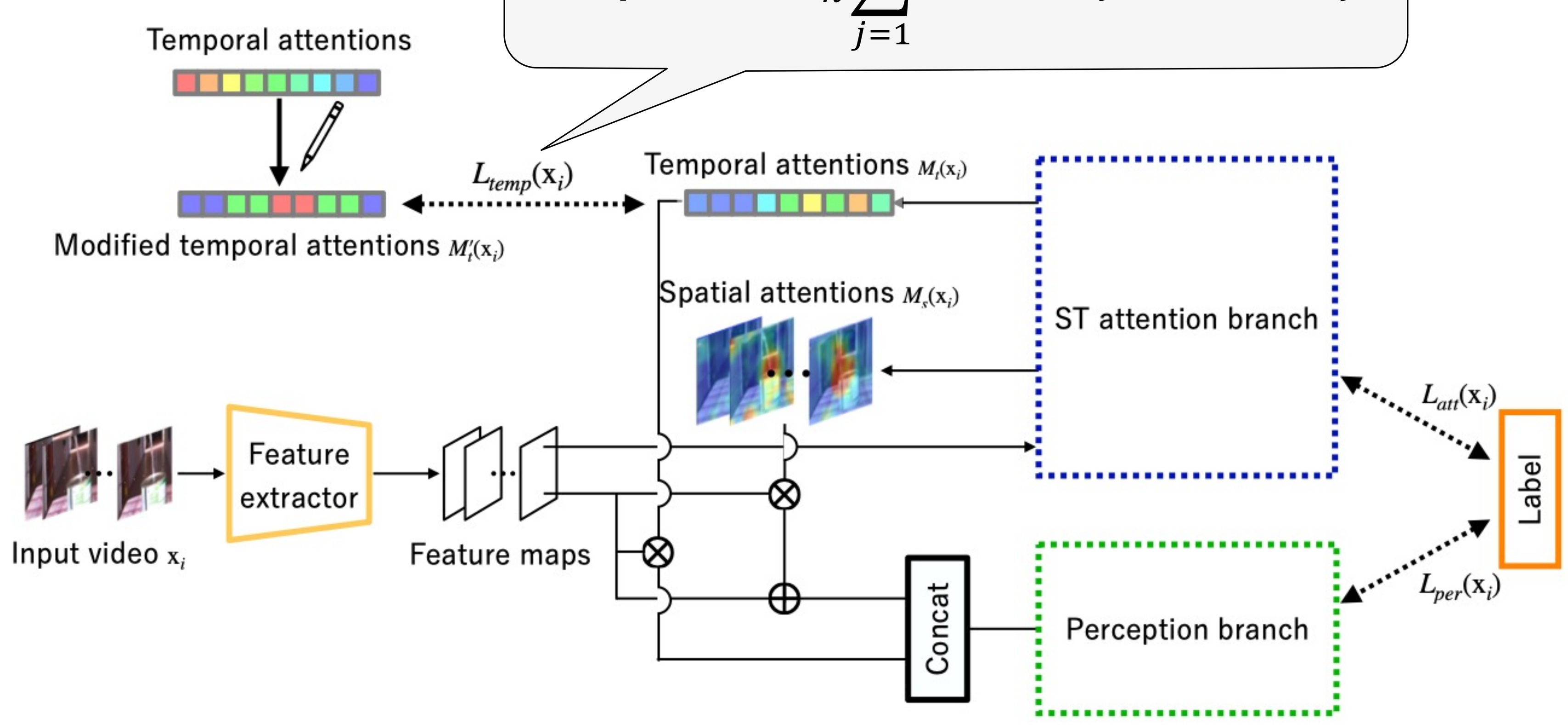
- Manually modify the temporal attentions collected in step 1
- Classify frames into three levels and edit temporal attentions



Step 3. Fine-tune the ST-ABN

- Fine-tune the branches of ST-ABN with modified temporal attentions
- We add a loss function L_{temp} to that of the ST-ABN
  - L_{temp} : Mean squared error with modified temporal attentions
- The ST-ABN optimizes its ST attention and perception branches.

$$L_{temp}(\mathrm{x}_i) = \gamma_t \frac{1}{n} \sum_{j=1}^{n} (\{M_t'(\mathrm{x}_i)\}_j - \{M_t(\mathrm{x}_i)\}_j)^2$$



$$L(\mathrm{x}_i) = L_{att}(\mathrm{x}_i) + L_{per}(\mathrm{x}_i) + L_{temp}(\mathrm{x}_i)$$

Experiments

- Experiment Details
  - Dataset : Something-Something v.2
  - Modified temporal attentions : 2,396 training samples in 8 action classes (1.5% of the total)

- Comparison with Conventional Models

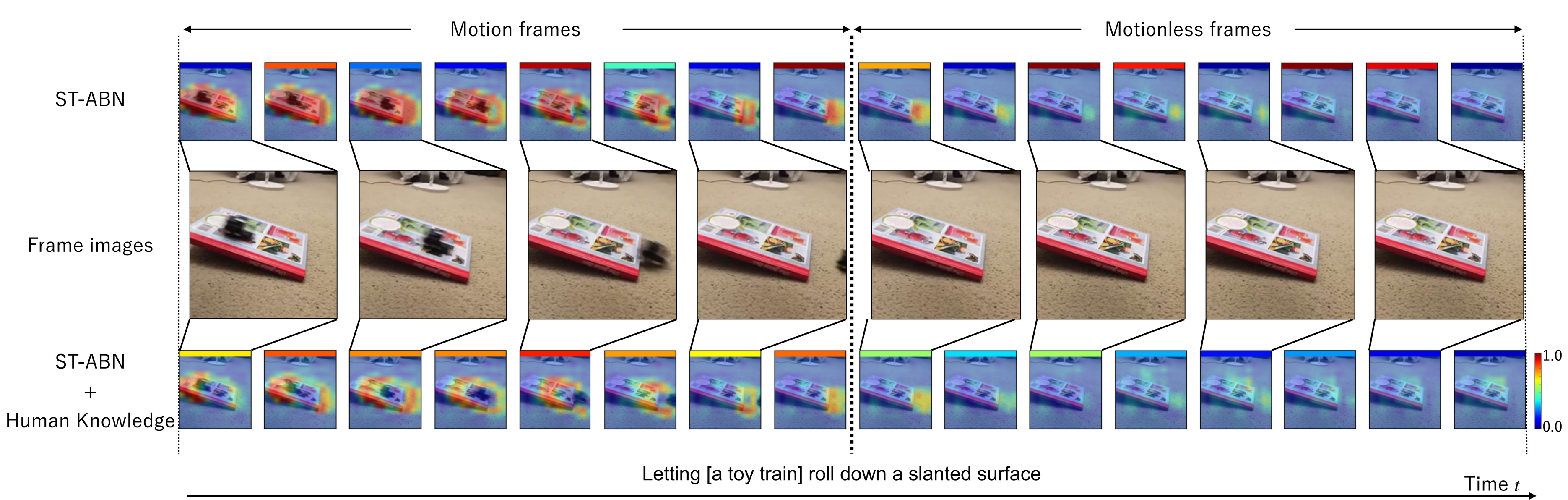
Method	Frames	Top-1 Acc.	Top-5 Acc.
3D ResNet-50	32	51.4	80.1
3D ResNet-50 + ST-ABN	32	**58.6**	**85.5**
3D ResNet-50	32 × 2	63.8	89.2
3D ResNet-50 + ST-ABN	32 × 2	**64.1**	**89.6**
3D ResNet-101	32	57.7	82.8
3D ResNet-101 + ST-ABN	32	**58.0**	**83.2**
3D ResNet-101	32 × 2	65.3	90.1
3D ResNet-101 + ST-ABN	32 × 2	**65.8**	**90.4**

Improved accuracy by introducing the ST-ABN into the backbone network

- Comparison with Embedding Human Knowledge (Backbone network : 3D ResNet-50)

Method	Modified classes	Other classes	All
ST-ABN	20.5	59.8	58.6
ST-ABN + Human Knowledge	**26.3**	**61.7**	**60.7**

Improved accuracy not only modified classes but also other classes



Embedding human knowledge into the ST-ABN obtain better attentions