# Temporal Attention with Conditional Inference for Large-Scale Multi-Label Video Classification

2018. Sep. 09

ECCV YouTube-8M Workshop

Team KANU

Eun-Sol Kim, Jongseok Kim, Kyoung-Woon On, Yu-Jung Heo,
Seong-Ho Choi, Hyun-Dong Lee and Byoung-Tak Zhang

kakaobrain & Biointelligence Lab.

# Submitted Model

# Experiments

| Criteria | Methods |
|---|---|
| Multimodal Inputs | Concatenate, Element-wise summation,<br>Attention, Differential Features |
| Temporal Aggregation | LSTM, GRU, Bidirectional LSTM, Hierarchical RNN<br>NetVLAD, CBHG |
| Classification Modules | Logistic Regression, Mixture of Experts,<br>Class Chaining, Conditional Inference |
| Additional Modules | Layer Normalization, Skip Connection, Dropout<br>Gradient Clipping |

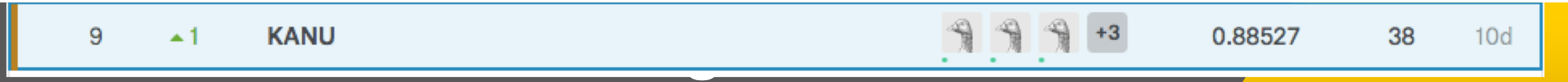
# Ensemble Models

- From large methods, six modes are selected by beam search under 1GB constraints
- The six models are combined with different weight values

# Ensemble Models

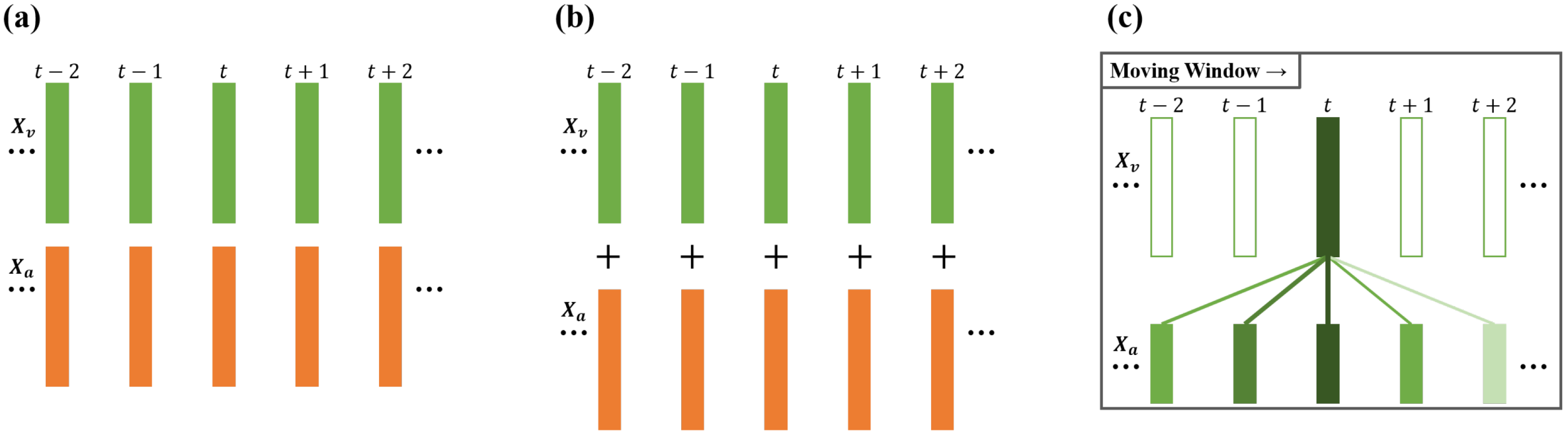
- Multimodal integ. (concat)-BLSTM-MoE2
- Multimodal integ. (att)-BLSTM-MoE2
- Multimodal integ. (concat)-BLSTM-CG-MoE2
- Multimodal integ. (concat)-NetVLAD-diff-C64-MoE4
- Multimodal integ. (sum)-NetVLAD-C64-MoE4
- Multimodal integ. (concat)-NetVLAD-C128-MoE4

- Size of the models: 162M, 163M,168M, 138M, 136M, and 200M
- Weight parameters: 0.2186, 0.2220, 0.1393, 0.1684, 0.1412, and 0.1102



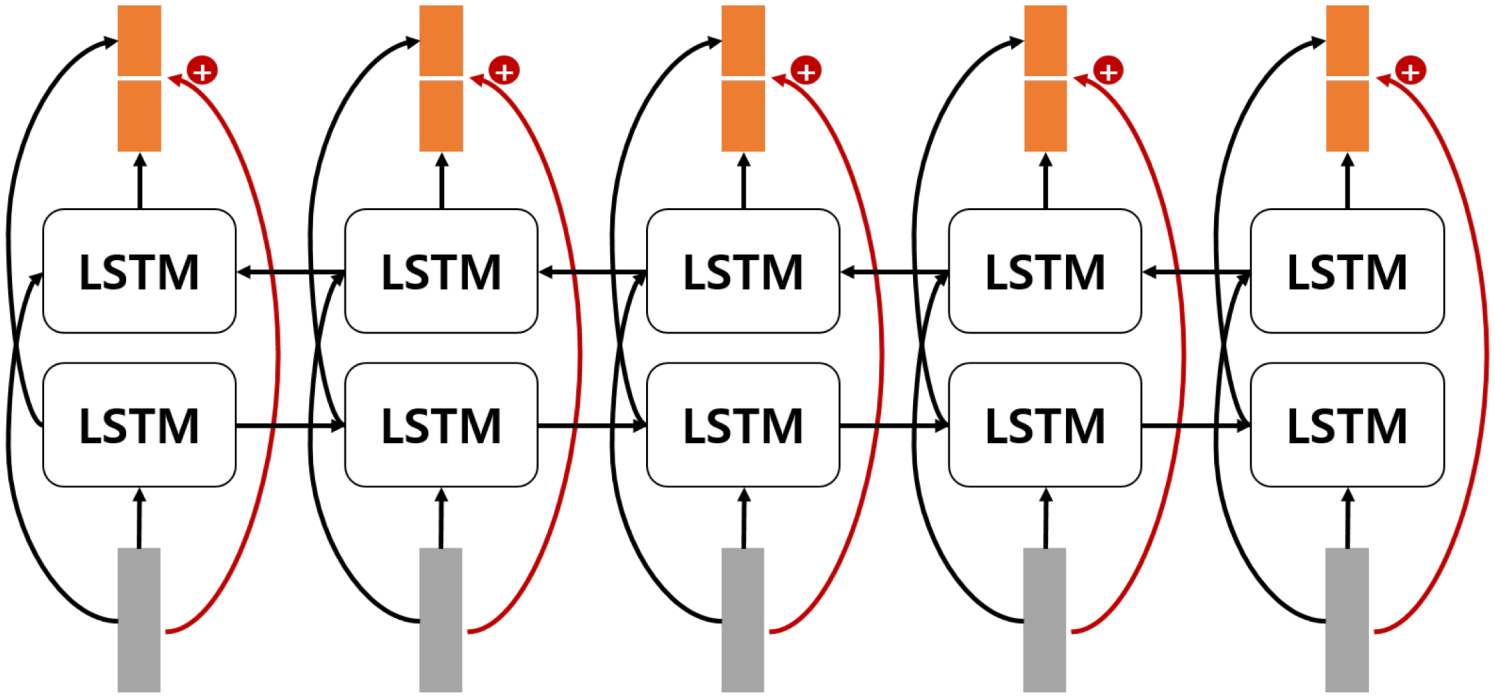
# Multimodal Inputs

- Multimodal Integrations
  - (a): Concatenation, (b): Summation, (c): video-guided temporal attention



# Temporal Aggregation

- Bidirectional LSTM with residual connection
- NetVLAD + context gating *




* A. Miech et al., 2017, Learnable pooling with Context Gating for video classification

# Problems on the YouTube-8M dataset

- Multimodal Representation Learning
  - Spatio-temporal Attention Methods
  - Temporal Structural Attention
  - Bi-linear (Compact) Pooling

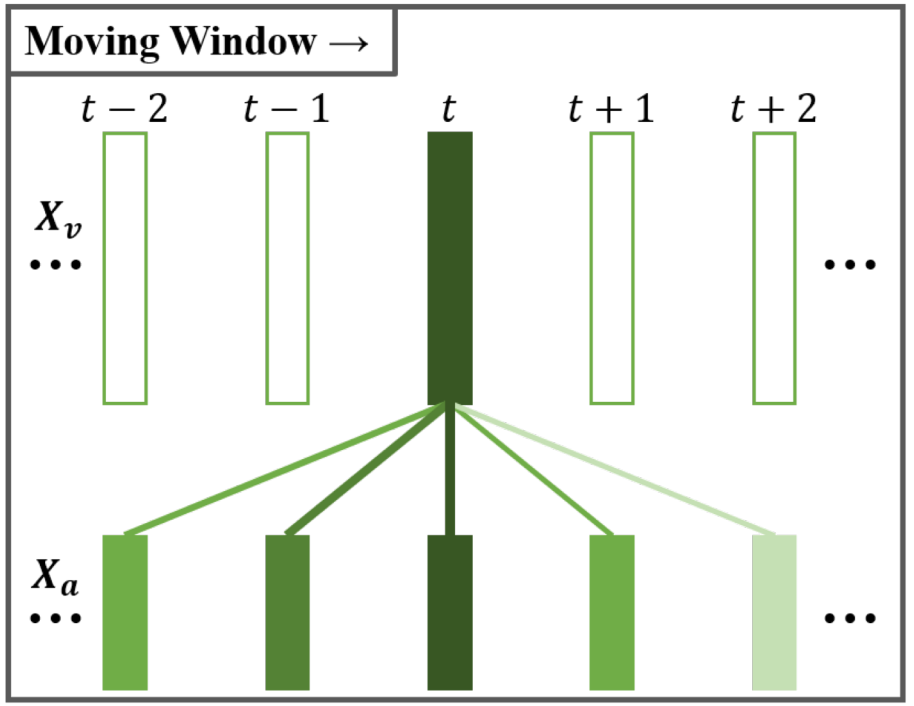
- Multi-label Classification
  - Class-chaining
  - Conditional Inference

# Multimodal Representation Learning

# Temporal attention on audio frames guided by a video frame

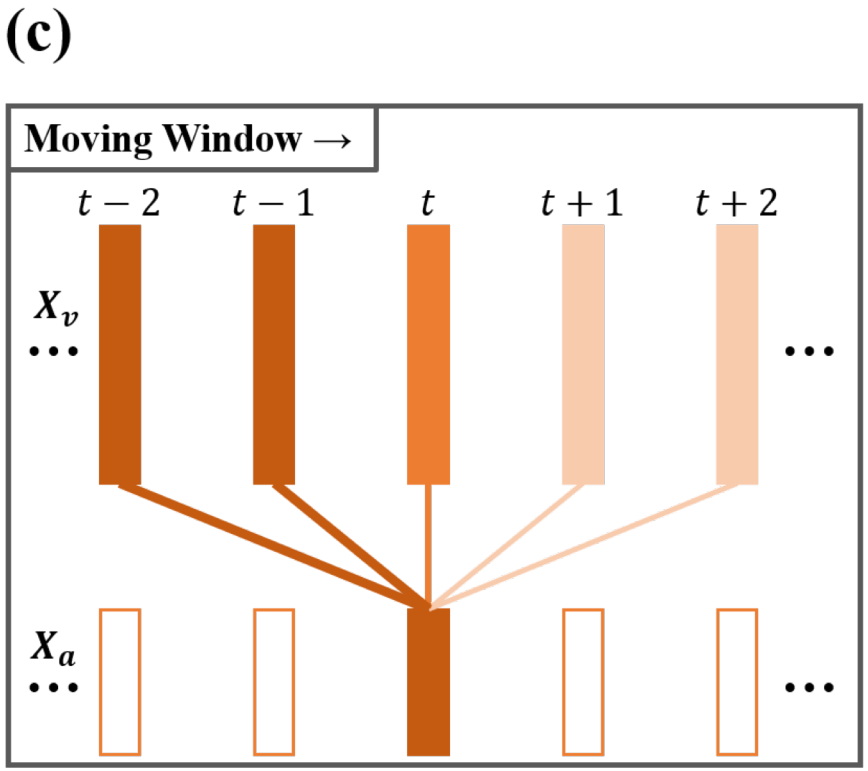
$$\mathbf{x}_f = \mathbf{x}_v + softmax(\mathbf{x}_v^\top \mathbf{W}_a^{att} \mathbf{X}_{a_{exp}}^{t-w:t}) \mathbf{X}_{a_{exp}}^{t-w:t}$$

**(b)**



# Temporal attention on video frames guided by an audio frame

$$\mathbf{x}_f = \mathbf{x}_{a_{exp}} + softmax(\mathbf{x}_{a_{exp}}^{\top} \mathbf{W}_v^{att} \mathbf{X}_v^{t-w:t}) \mathbf{X}_v^{t-w:t}$$



# Experimental Results

| Attention Method | Window Size $w$ | Accuracy (GAP) |
|---|---|---|
| None | None | 0.858 |
| Image Guided Attention | 5 | 0.86071 |
| Image Guided Attention | 9 | 0.86078 |
| Image Guided Attention | 13 | 0.85920 |
| Image Guided Attention | all | 0.86129 |
| Audio Guided Attention | 5 | 0.85670 |

+ BLSTM, MoE (#4)

# Multi-label Classification

# Conventional Methods on MLC

Given L classes
Predict M multiple Classes

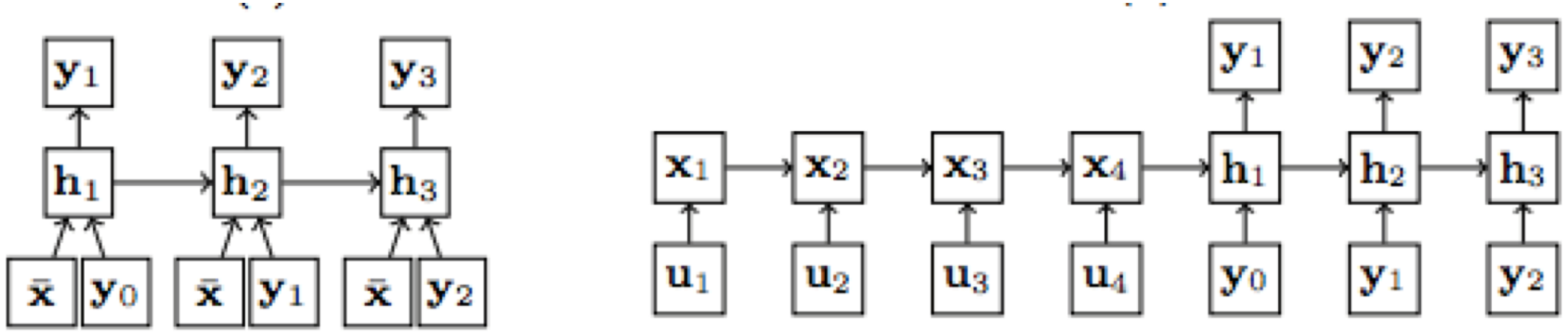
- Binary Relevance
  - Decomposing MLC problem into a number of independent binary task
  - (-) hard to learn the dependency and correlations among the labels
- Label Powerset(LP)
  - Redefine the label powerset ($2^L$), and one-class classification
  - (-) Limited training examples / Cannot predict unseen label sets
- Class-chaining
  - Sequentially predict label sets
  - (-) label ordering problem

# Class Chaining with RNN

- Sequentially predict $M$ labels
- Factorize the conditional probability

$$p(\mathbf{y}|\mathbf{x}) = \prod_{I=1}^{M} p(y_i|\mathbf{x}, \mathbf{y}_{<I})$$

  - Calculated in a chaining manner, implement using RNN
  - Use the previous predicted labels as features



# Conditional Inference

- Instead of ordering labels, build the function to learn mapping to $y$ given $x$ and any combination of previously observed labels
  - $p(\boldsymbol{y}|x) = \prod_{i=1}^{q} f\ (x, y_{<\phi_i})$
  - Previously observed labels is $q$-dimensional vector $\{0, 1\}^q$ where observed as positive set to 1 and others 0
  - Constraint:
    - Forcing previously observed labels to the positive ones only.
  - Iteratively perform as far as it doesn't seriously impair constraint

# Conditional Inference

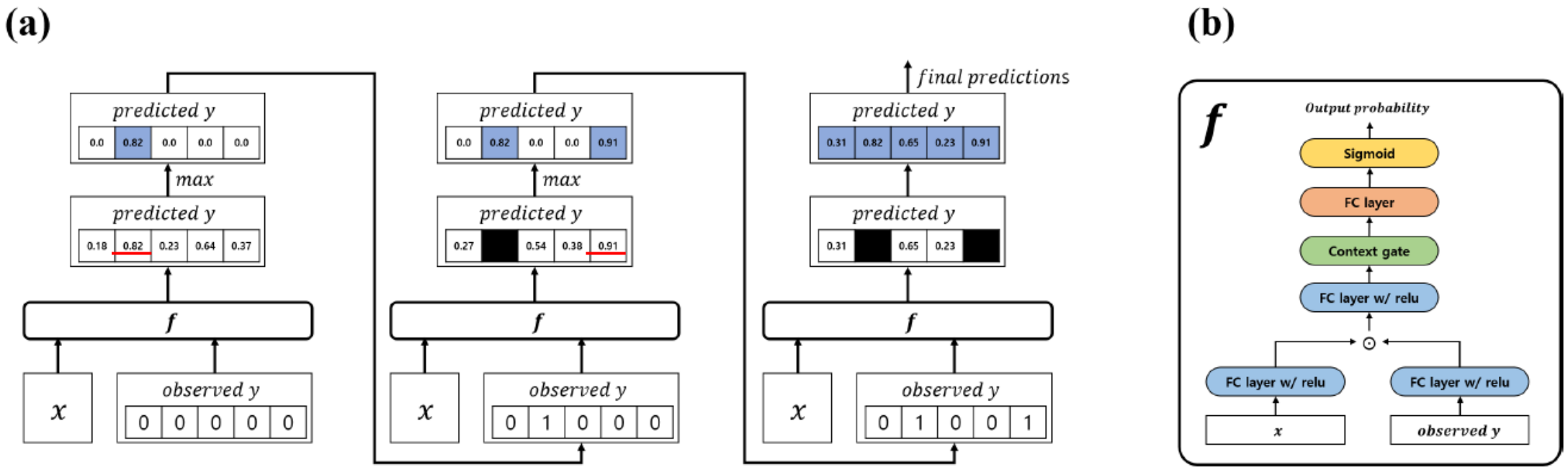



**Fig. 3.** (a): An illustration of the conditional inference procedure on 5-labels and 2-steps situation. (b): Core neural network architecture of conditional inference.

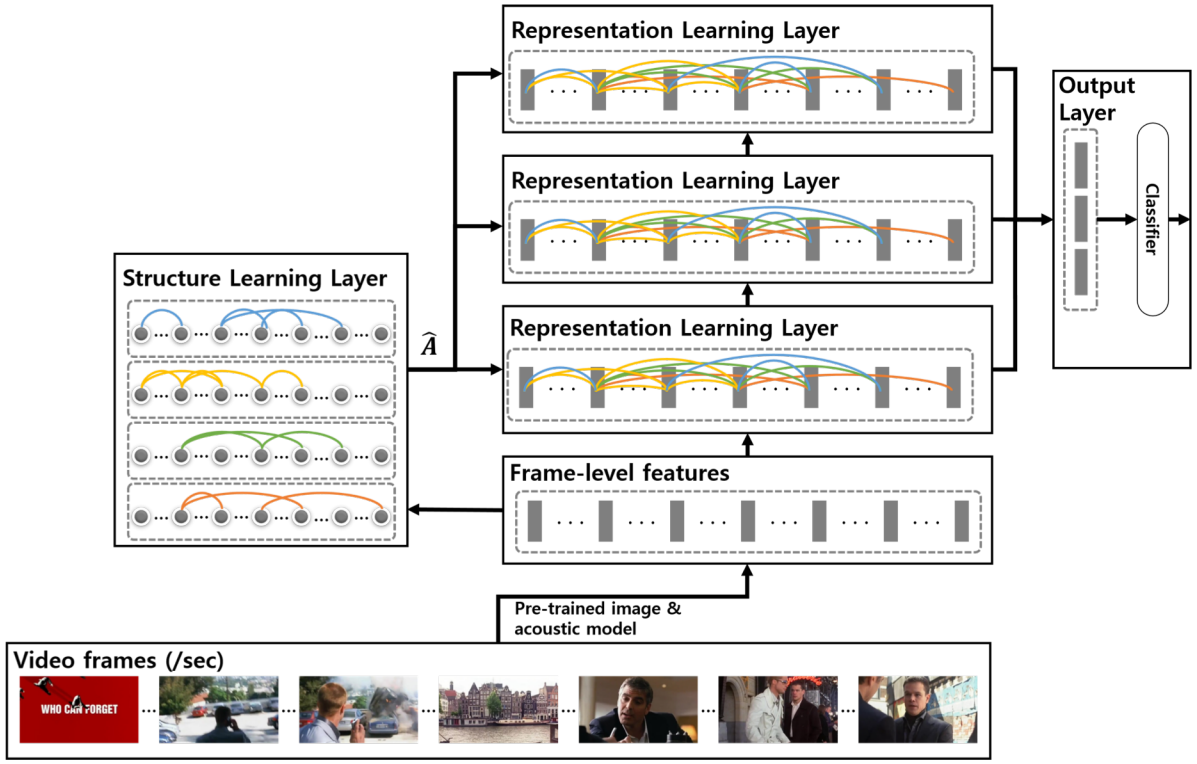
# Conditional Inference

| MLC Method | Structure | GAP |
|---|---|---|
| Logistic Model | | 0.7942 |
| Class Chaining | GRU | 0.795 |
| Mixture of Experts | (# of Experts) 2 | 0.8282 |
| | (# of Experts) 3 | 0.8296 |
| | (# of Experts) 4 | 0.8305 |
| Conditional Inference | (# of Steps) 1 | 0.8385 |
| | (# of Steps) 2 | 0.8398 |
| | (# of Steps) 3 | 0.8407 |
| | (# of Steps) 4 | 0.8410 |
| | (# of Steps) 5 | 0.8403 |

# Further Works

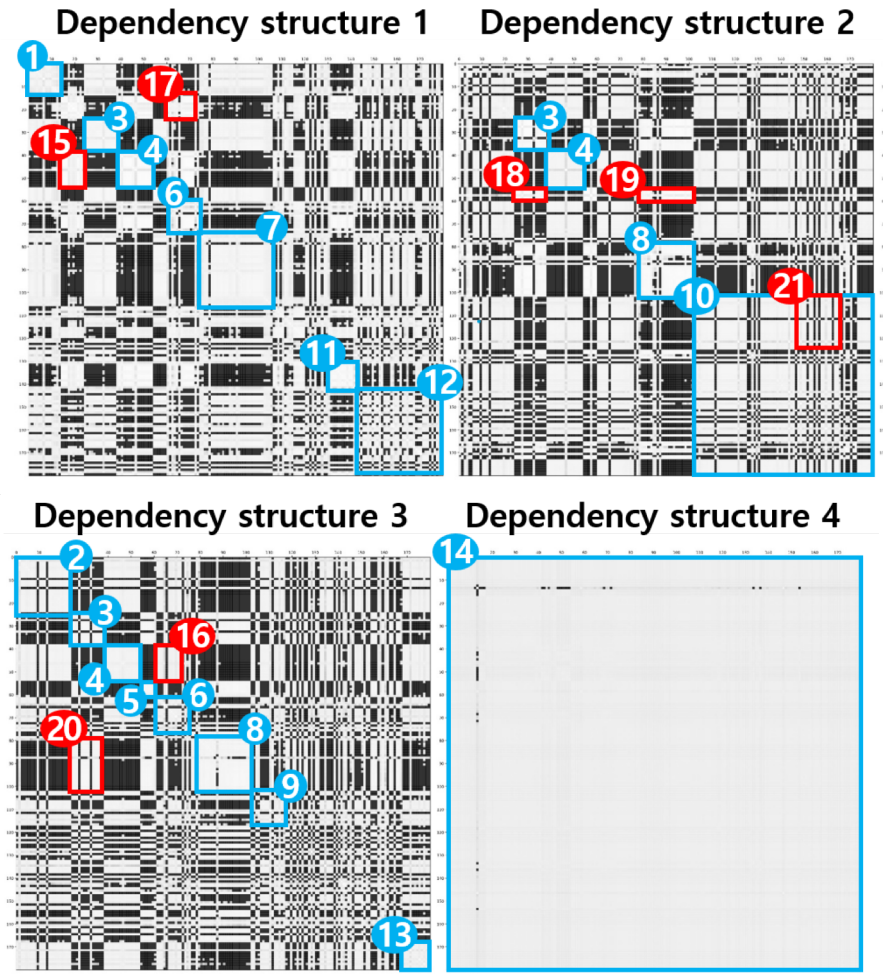
# Temporal Dependency Networks

- Learning video data by discovering temporal dependency structure
  - represents video data as a temporal graph
    - Node:  frame of a video
    - Edge: the dependency between two frames
  - Structure learning
  - Representation learning



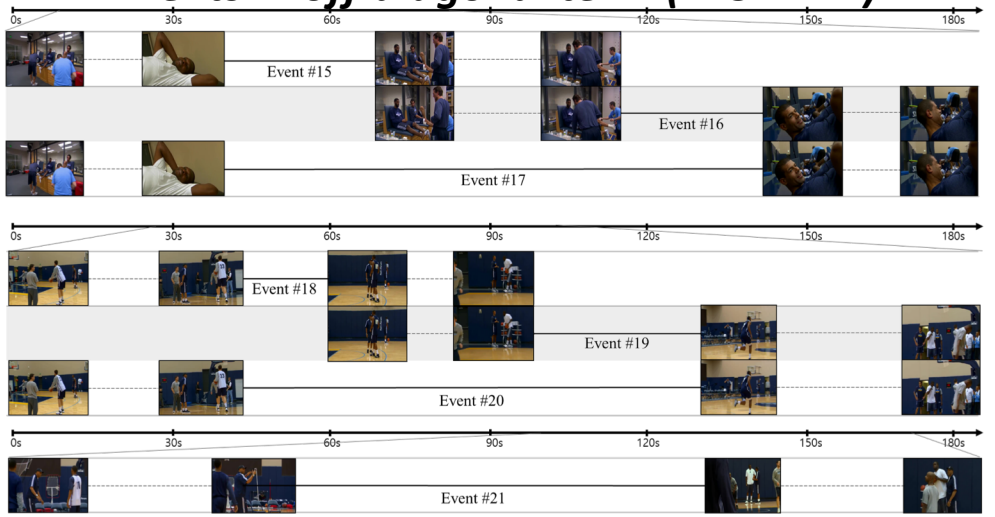
# Temporal Dependency Networks

***Learned adjacency matrices***



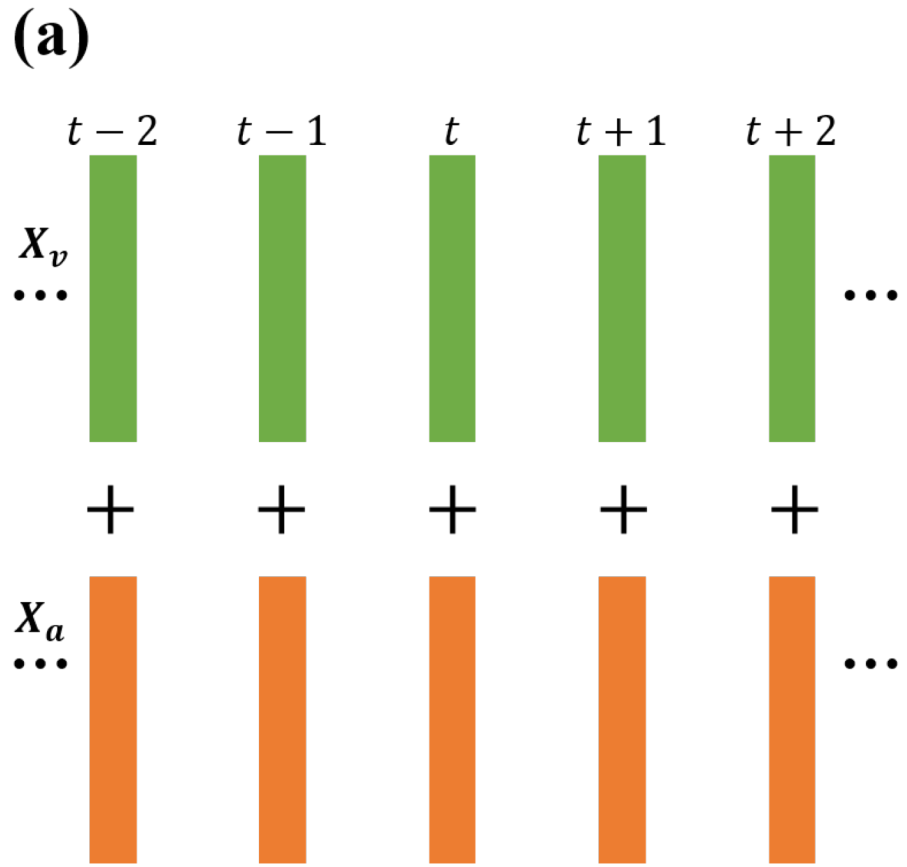

***Events in diagonal term (#7 - #14)***


***Events in off-diagonal term (#15 - #21)***



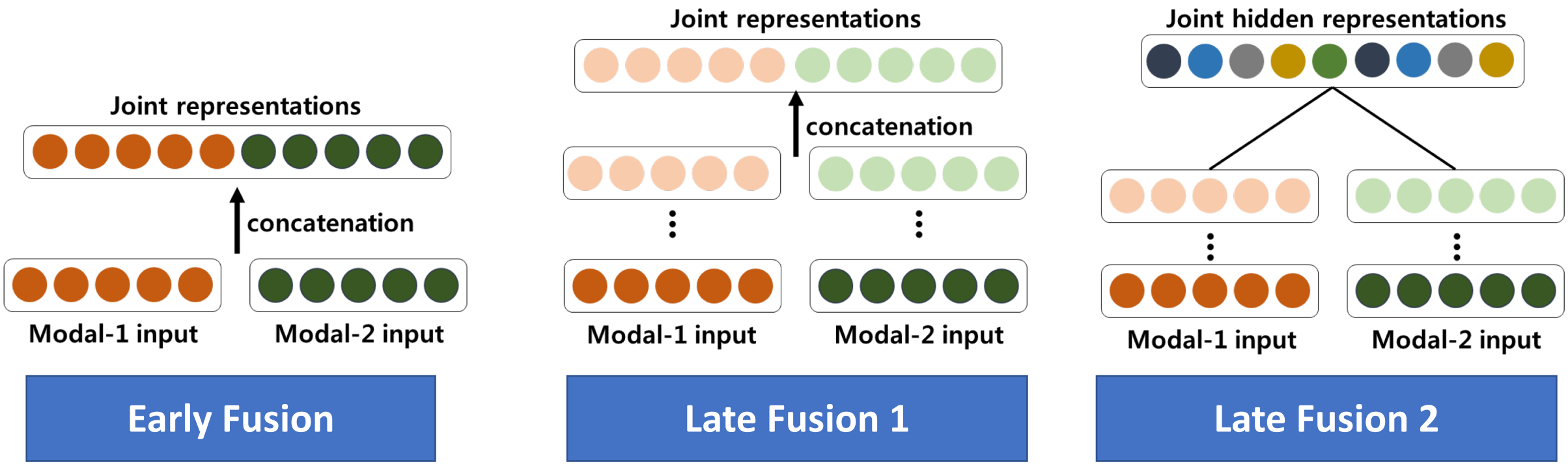
# Appendix

# Element-wise summation after a linear transformation




$$\mathbf{x}_{a_{exp}} = \mathbf{W}_{va}\mathbf{x}_a + \mathbf{b}_{va}$$
$$\mathbf{x}_f = \mathbf{x}_v + \mathbf{x}_{a_{exp}}$$

# Conventional Methods



# Multi-label Classification(MLC)

- Training data: $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_1^n, \boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^D, \boldsymbol{y}_i \in \{0,1\}^L$
- Learning a mapping: $g: \mathcal{X} \to \{0,1\}^L$
- Each instance $\boldsymbol{x}_i$ is associated with a set of relevant labels $\boldsymbol{y}_i$

# (cf) Multiclass Classification

- Training data $\{(\boldsymbol{x}_i, y_i)\}_1^n, \boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^D, y_i \in \mathcal{Y} = \{1,2,\dots,L\}$
- Learn a mapping $f: \mathcal{X} \to \mathcal{Y}$
- Each instance $\boldsymbol{x}_i$ is associated with a single relevant label $y_i$