Supplement Materials for “Weakly-Supervised Action Localization by Hierarchically-structured Latent Attention Modeling”

Guiqin Wang[1†] Peng Zhao[1] Cong Zhao[3,4] Shusen Yang[3,4] Jie Cheng[2] Luziwei Leng[2]
Jianxing Liao[2] Qinghai Guo[2*]
[1] School of Computer Science and Technology, Xi’an Jiao Tong University
[2] ACS Lab, Huawei Technologies
[3] School of Mathematics and Statistics, Xi’an Jiao Tong University
[4] National Engineering Laboratory for Big Data Analytics, Xi’an Jiao tong University

1. Introduction

This supplementary material contains four parts:

- Section 2 provides the derivation of the ELBO loss function.
- Section 3 provides a detailed explanation table of our paper’s parameters.
- Section 4 shows more detailed description for architecture of our model.
- Section 5 describes the inference process of our attention-based classification module(EFC) and its interaction with our change-points detection module(DFC).
- Section 6 provides a visualisation sample of the Change-point Module to show the effectiveness of our proposed model.

2. Derivation of ELBO

$$\mathcal{L}_{ELBO} = \sum_{t=1}^{T} \mathbb{E}_{q(v_t^{1,2})}[\log p(x_t|v_t^{1,2})] - \sum_{n=1}^{2}\sum_{t=1}^{T} \mathbb{E}_{q(v_t^{>n}, v_{<t}^{n})}[D_{KL}(q_\phi(v_t^n|x_t, v_t^{>n}, v_{<t}^n)||p_\theta(v_t^n|v_t^{>n}, v_{<t}^n))], \tag{1}$$

As mentioned in the paper, we aim to estimate the true posterior distribution $p(v_{1:T}^1, v_{1:T'}^2|x_{1:T})$ through the approximation distribution $q(v_{1:T}^1, v_{1:T'}^2|x_{1:T})$. (in the following, in case of no ambiguity, we use $p(v|x)$ to denote the true posterior, and use $q(v)$ to denote the approximation for simplicity). The optimal $q(v)$ is then taken as

$$q^*(v) = \underset{q(v)}{\arg\min}\, D_{KL}(q(v)||p(v|x)). \tag{2}$$

To compute the KL-divergence, we have

$$\begin{aligned} D_{KL}(q(v)||p(v|x)) &= -\int_v q(v)\log\left[\frac{p(v|x)}{q(v)}\right] dv \\ &= \int_v q(v)\log q(v)dv - \int_v q(v)\log p(v|x)dv \\ &= \mathbb{E}_q[\log q(v)] - \mathbb{E}_q[\log p(v|x)] \\ &= \mathbb{E}_q[\log q(v)] - \mathbb{E}_q[\log p(v,x)] + \mathbb{E}_q[\log p(x)], \end{aligned} \tag{3}$$

where in the last equation we use the conditional probability. Note that in the third term $\log p(x)$ is independent to the distribution q, hence we have

$$\begin{aligned} D_{KL}(q(v)||p(v|x)) = \mathbb{E}_q[\log q(v)] - \mathbb{E}_q[\log p(v,x)] \\ + \log p(x). \end{aligned} \tag{4}$$

The ELBO is then defined as

$$\begin{aligned} \mathcal{L}_{ELBO} &:= \mathbb{E}_q[\log p(v,x)] - \mathbb{E}_q[\log q(v)] \\ &= \log p(x) - D_{KL}(q(v)||p(v|x)). \end{aligned} \tag{5}$$

Minimizing $D_{KL}(q(v)||p(v|x))$ then becomes maximizing $\mathcal{L}_{ELBO}$ since $\log p(x)$ is fixed. In specific, for $\mathcal{L}_{ELBO}$, we

[†] Work is done during the internship at ACS Lab, Huawei Technologies.
[*] Corresponding author (guoqinghai@huawei.com).

Table 1. List of Parameter Explanation in our Paper

Parameter	Explanation
s	the start boundaries of action instance in the input video
e	the end boundaries of action instance in the input video
c	the class prediction of action instance
q	the confidence score of action instance
I	the segment sample of input video
P	the change-point of input video
X	the extraction feature of input video
M	the activation score of class for feature
T	the length of input video
T'	the num of change-points
u	the decoder output from top-down
d	the output of transition model
v	the random variable representing the hidden state
$p(\cdot)$	the prior model
$q(\cdot)$	the posterior model
ch	the assumption that a change-point exists
st	the assumption that the current state stays static

have

$$\begin{aligned}\mathcal{L}_{ELBO} &= \mathbb{E}_q[\log p(x|v)p(v)] - \mathbb{E}_q[\log q(v)] \\ &= \mathbb{E}_q[\log p(x|v)] + \mathbb{E}_q[\log p(v)] - \mathbb{E}_q[\log q(v)] \\ &= \mathbb{E}_q[\log p(x|v)] + \mathbb{E}_q[\frac{\log p(v)}{\log q(v)}] \\ &= \mathbb{E}_q[\log p(x|v)] + \int_q q(v)\frac{\log p(v)}{\log q(v)}dv \\ &= \mathbb{E}_q[\log p(x|v)] - D_{KL}(q(v)||p(v)) \\ &= \sum_{t=1}^{T}\mathbb{E}_{q(v_t^{1,2})}[\log p(x_t|v_t^{1,2})] - \sum_{n=1}^{2}\sum_{t=1}^{T}\end{aligned}$$

$$\mathbb{E}_{q(v_t^{>n},v_{<t}^{n})}[D_{KL}(q_\phi(v_t^n|x_t, v_t^{>n}, v_{<t}^n)||p_\theta(v_t^n|v_t^{>n}, v_{<t}^n))]. \tag{6}$$

Note that in the last equation, we use the fact that the distribution of hidden state v_t^n depends on the past states in the same level $v_{<t}^n$ and the states in higher level $v_t^{>n}$, and similar dependency exists for posterior distribution of v_t^n.

3. Parameter List

In this section, we summarize the specific meaning of each parameter, which appears in the paper, following the table1 that shows detail.

4. Model Architecture

Our model mainly consists of two parts(Attention-based Classification Module and Change-point detection Module).

The Attention-based Classification Module consists of three branches, each branch consists of a Conv1D and a softmax layer. Formally, we denote the branch attention module as following:

$$output = Softmax(Conv1D(X, \theta_{att})), \tag{7}$$

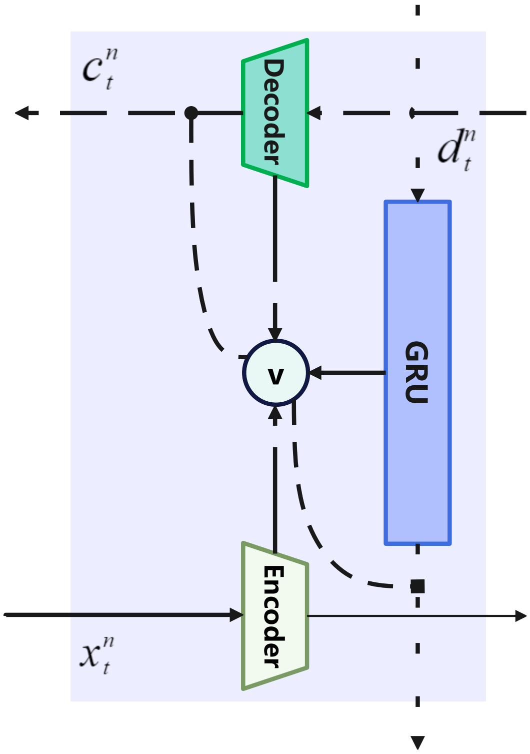

Figure 1. =Change-point Detection Module block architecture. Different information channels use different model architecture: Encoder, Decoder, Temporal Model.

where θ_{att} denotes the trainable convolution layer parameters of the attention branch, which is an FC layer.

The Change-point Detection Module is made up of $T \times 2$ detection blocks, where T is the length of input video, 2 denotes the 2-level. As Figure1 shows, each detection block consists of three parts (*i.e.*, Encoder, Decoder, Temporal Model). As Table 2 shows, Encoder is a combination of a fully-connected(FC) layer(1024-d), which is used for VAE, and an FC layer(64-d), which is passed into the latent model; Decoder is a combination of an FC layer(1024-d), which is used for VAE, and an FC layer(64-d), which is passed into the latent model; Temporal model is a recurrent GRU model(256-d), and an FC layer (64-d), which is passed into the latent model; Latent model, aiming at generating v, uses an FC network and parameterise a diagonal Gaussian, with the output dimension of 2×64.

5. Details of EFC

For attention-based foreground classification(EFC) module, we set the result of attention-based classification module [2] as our baseline. Based on the the baseline, we select the foreground change-points(adjacent to baseline) as

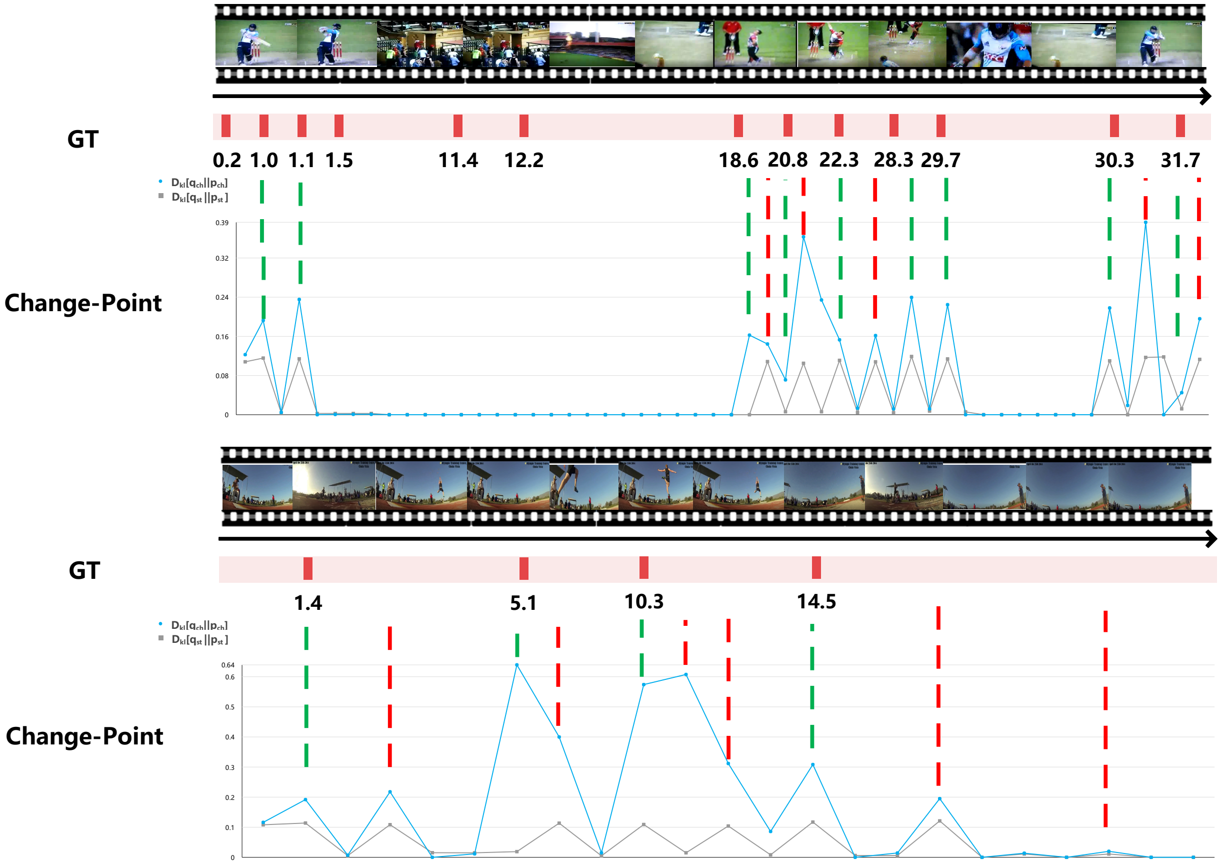

Figure 2. Visualization of qualitative results on the THUMOS-14 dataset. From the above qualitative results, we can conclude that our proposed Change-point Detection Module is greatly beneficial to locate action boundaries and help us achieve more precise temporal action localization results. The green dashed line is the correct boundary and red dashed one is wrong. As the fold line shows, in line with the described detection criteria in the submitted paper, if one of the gray lines falls below the blue line, a change-point is considered to be detected.

Table 2. The architecture of our model.

Name	Layer	Input Size	Output Size
Input Clip	-	-	$T \times 2048$
Encoder	FC layer	Input Clip	$T \times 1024$
	FC layer	$T \times 1024$	$T \times 64$
Decoder	FC layer	$T \times 1024$	$T \times 1024$
	FC layer	$T \times 1024$	$T \times 64$
Temporal Model	GRU	$T \times 256$	$T \times 256$
	FC layer	$T \times 256$	$T \times 64$
Latent Model	FC layer	-	$T \times (2 \times 64)$

action boundaries. Meanwhile, we utilize the longest common sub-sequence(LCS) [1] to delete the redundant adjacent change-points during the inference process. In specific, for two adjacent change-points A and B, we construct two snippets $l_{AC} = \{n_1, n_2, ..., n_a\}$ and $l_{BC} = \{m_1, m_2, ..., m_b\}$ by connecting A and B with a third change-point C. We calculate the cosine similarity of snippets n_i and m_j:

$$cos_{i,j} = cos(n_i, m_j). \tag{8}$$

We set the similarity threshold as 0.65. If $cos > 0.65$, then the calculated common sub-sequence is extended by that element and thus $L(i,j) = L(i-1, j-1)+1$. If $cos \leq 0.65$, the largest length calculated before is retained for $L(i,j)$:

$$L(i,j) = \begin{cases} L(i-1, j-1) + 1 & , cos_{i,j} > 0.65 \\ max\{L(i-1,j), L(i,j-1)\}, & cos_{i,j} \leq 0.65 \end{cases} \tag{9}$$

where $L(i,j) = 0$ if $i = 0$ or $j = 0$.

Finally, we compare the resulted $L = L(a,b)$ with the baseline(the boundaries result of attention-based classification module), we delete the the redundant change-points to obtain action boundaries. Specifically, we compare L with the length K of selected baseline(the most overlapped snippet with l_{BC} from baseline snippets). If $L < K/2$ (which indicates that l_{AC} and l_{BC} represent different actions), we retain the change-point A and the change-point B, otherwise we delete A.

6. Visualisation Result

Figure 2 shows two examples comparing the ground truth to prove effectiveness of our change-point detection module. Our model is able to produce precise boundaries and, consequently, better action localization through the change-point detection module. For action localization task, our model is able to obtain more accurate temporal boundaries. This leads to average precision improvements for higher IoU overlap thresholds. However, since we select the right candidates through foreground and background detection, this relies on the attention model to filter boundaries, which limits the performance of our model.

References

[1] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009, 2022.

[2] Sanqing Qu, Guang Chen, Zhijun Li, Lijun Zhang, Fan Lu, and Alois Knoll. Acm-net: Action context modeling network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2104.02967*, 2021.