Weakly Supervised Energy-Based Learning for Action Segmentation

Jun Li, Peng Lei, Sinisa Todorovic

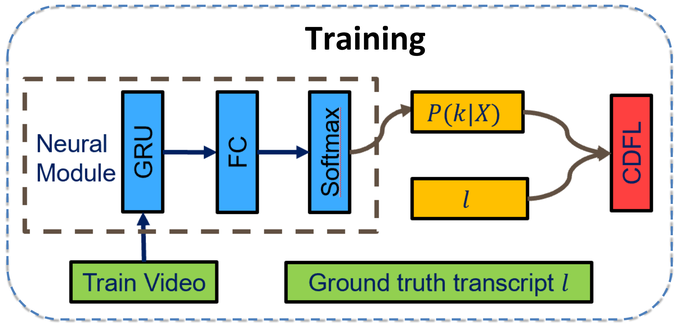
ICCV 2019 Seoul, Korea

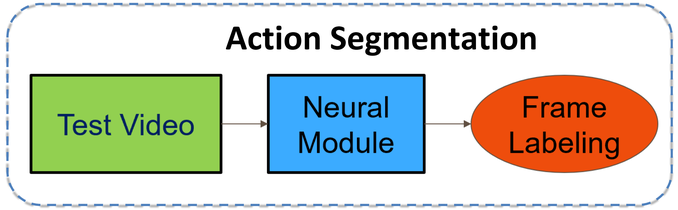
Problem:

Predict frame labels, when the ground truth in training is limited and specifies only the temporal ordering of actions, instead of their temporal extents.

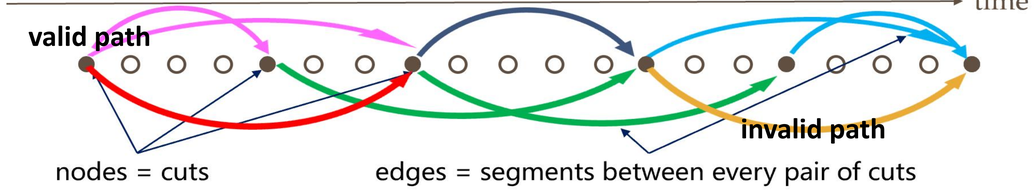
Key ideas for weakly-supervised training:

- Train a frame labeler on an energy-based loss which accounts for all valid and invalid video segmentations.
- But there are exponentially many valid segmentations.
- Use the logadd to efficiently compute the loss.





Transcript: $(a_1 \rightarrow a_2 \rightarrow a_3)$ = (pour_cereals → pour_milk → stir_cereals)



Edge weight: $w_{ii'}(a) = \sum_{t \in (i,i')} -\log p(a|x_t)$ ← GRU's softmax score for action a at frame t

Path energy:

$E_\pi = \sum_{e_{ii'} \in \pi} w_{ii'}(a^\pi_{ii'})$ ← Sum of edge weights in the path

Energy of all valid paths: $L_{\mathrm{F}} = -\log\left(\sum_{\pi \in \mathcal{P}^V} \exp(-E_\pi)\right)$

Exponentially many paths

Efficient loss computation:

$$\ell_{i'}(\boldsymbol{a}_{1:n}) = \mathrm{logadd}(\{\ell_i(\boldsymbol{a}_{1:n-1}) + w_{ii'}(a_n) : i < i'\})$$

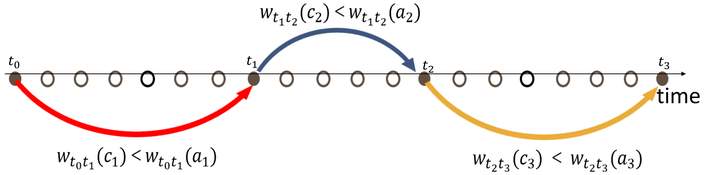
where: $\mathrm{logadd}(a,b) = -\log(\exp(-a) + \exp(-b))$

Constrained discriminative forward loss (CDFL):

$$L_{\mathrm{CDF}} = \mathrm{logadd}(\mathcal{P}^V) - \mathrm{logadd}(\mathcal{P}^{I_c})$$

a subset of hard invalid paths whose edge scores are lower than those of true actions.

An example of hard invalid path:



Results:

Breakfast	Mean accuracy over frames(%)
OCDC(Bojanowski et al. ECCV 2014)	8.9
CTC(Huang et al. ECCV 2016)	21.8
HTK(Kuehne et al. WACV 2016)	25.9
ECTC(Huang et al. ECCV 2016)	27.7
HMM/RNN(Richard et al. CVPR 2017)	33.3
TCFPN(Ding et al. CVPR 2018)	38.4
NN-Viterbi(Richard et al. CVPR 2018)	43.0
D3TW(Chang et al. CVPR 2019)	45.7
Our CDFL	**50.2**

ACKNOWLEDGMENT: DARPA XAI Award N66001-17-2-4029 and AFRL STTR AF18B-T002.