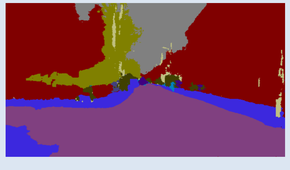
Budget-Aware Deep Semantic Video Segmentation

Behrooz Mahasseni, Sinisa Todorvic and Alan Fern

Problem: Semantic Video Segmentation

- Goal : Label every pixel in every frame



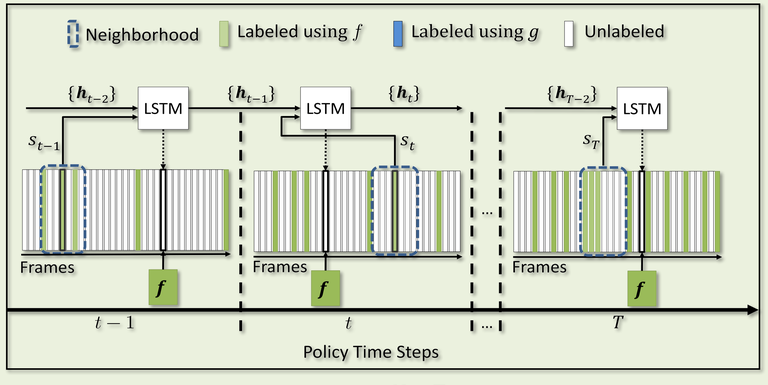
State of the art:

- Per frame runtime : 160ms ~ 1450ms
- Varying time budgets poorly studied

Key Ideas:

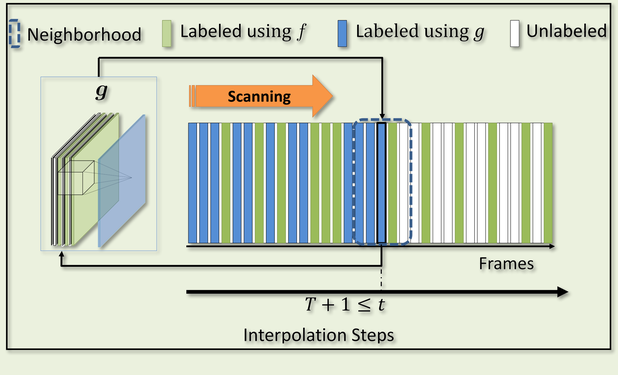
- "Black-box" frame-based video segmentation
- Budget-aware policy for frame selection
- Label interpolation in the remainder

Policy for Frame Selection



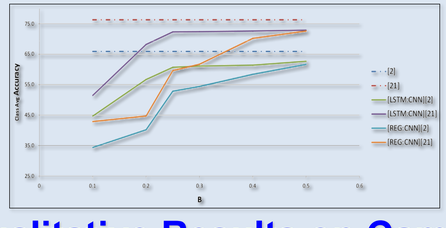
Approximate Gradient: $\nabla_{\theta_\pi} J \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T} \left[\nabla_{\theta_\pi} \log \pi(l_t^n | \boldsymbol{h}_{t-1}^n, o_t^n) R_t(\boldsymbol{h}_t^n) \right]$

Pixel-wise Label Interpolation

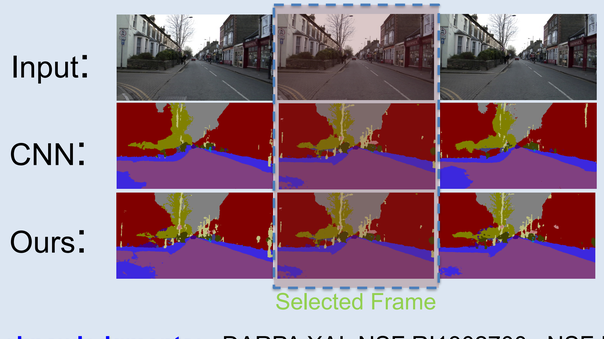


Budget vs Accuracy in [%] on CamVid

Method	Time for π	Time for g	Time for f	Class Avg	Mean I/U
REG+CNN (B=0.25 × B_{max})	0	184.5	2928.4	62.3	49.3
REG+CNN (B=0.5 × B_{max})	0	97.3	4170.8	72.6	59.6
LSTM + CNN (B=0.25 × B_{max})	20.4	196.8	2934.7	72.3	59.4
LSTM+CNN (B=0.5 × B_{max})	8.8	113.9	4181.3	73.3	59.8



Qualitative Results on CamVid



Acknowledgments: DARPA XAI, NSF RI1302700 , NSF IIS1619433.